University of Co-operative and Management, Thanlyin

Department of Statistics

BASIC STATISTICS
MAR 112

FIRST, QUARTER

UNIVERSITY OF CO-OPERATIVE AND MANAGEMENT, THANLYIN
DEPARTMENT OF STATISTICS

BASIC STATISTICS
MAR 112

FIRST QUARTER

Contents

Chapter 1	The Nature of Probability and Statistics	1
Chapter 2	Frequency Distributions and Graphs	13
Chapter 3	Data Description	65
Chapter 4	Probability and Counting Rules	127
Chapter 5	Correlation and Regression	187

1

The Nature of Probability and Statistics

© Shutterstock/Monkey Business Images RF

STATISTICS TODAY

Is Higher Education "Going Digital"?

Today many students take college courses online and use eBooks. Also, many students use a laptop, smartphone, or computer tablet in the classroom. With the increased use of technology, some questions about the effectiveness of this technology have been raised. For example,

How many colleges and universities offer online courses?

Do students feel that the online courses are equal in value to the traditional classroom presentations?

Approximately how many students take online courses now?

Will the number of students who take online courses increase in the future?

Has plagiarism increased since the advent of computers and the Internet?

Do laptops, smartphones, and tablets belong in the classroom?

Have colleges established any guidelines for the use of laptops, smartphones, and tablets?

To answer these questions, Pew Research Center conducted a study of college graduates and college presidents in 2011. The procedures they used and the results of the study are explained in this chapter. See Statistics Today—Revisited at the end of the chapter.

OUTLINE

Introduction

1–1 Descriptive and Inferential Statistics

1–2 Variables and Types of Data

1–3 Data Collection and Sampling Techniques

1–4 Experimental Design

1–5 Computers and Calculators

Summary

OBJECTIVES

After competing this chapter, you should be able to:

1. Demonstrate knowledge of statistical terms.
2. Differentiate between the two branches of statistics.
3. Identify types of data.
4. Identify the measurement level for each variable.
5. Identify the four basic sampling techniques.
6. Explain the difference between an observational and an experimental study.
7. Explain how statistics can be used and misused.
8. Explain the importance of computers and calculators in statistics.

Introduction

You may be familiar with probability and statistics through radio, television, newspapers, and magazines. For example, you may have read statements like the following found in newspapers.

- A recent survey found that 76% of the respondents said that they lied regularly to their friends.
- *The Tribune Review* reported that the average hospital stay for circulatory system ailments was 4.7 days and the average of the charges per stay was $52,574.
- Equifax reported that the total amount of credit card debt for a recent year was $642 billion.
- A report conducted by the SAS Holiday Shopping Styles stated that the average holiday shopper buys gifts for 13 people.
- The U.S. Department of Agriculture reported that a 5-foot 10-inch person who weighs 154 pounds will burn 330 calories for 1 hour of dancing.
- The U.S. Department of Defense reported for a recent year that the average age of active enlisted personnel was 27.4 years.

Unusual Stats

Of people in the United States, 14% said that they feel happiest in June, and 14% said that they feel happiest in December.

Statistics is used in almost all fields of human endeavor. In sports, for example, a statistician may keep records of the number of yards a running back gains during a football game, or the number of hits a baseball player gets in a season. In other areas, such as public health, an administrator might be concerned with the number of residents who contract a new strain of flu virus during a certain year. In education, a researcher might want to know if new methods of teaching are better than old ones. These are only a few examples of how statistics can be used in various occupations.

Interesting Fact

Every day in the United States about 120 golfers claim that they made a hole-in-one.

Furthermore, statistics is used to analyze the results of surveys and as a tool in scientific research to make decisions based on controlled experiments. Other uses of statistics include operations research, quality control, estimation, and prediction.

Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

Historical Note

A Scottish landowner and president of the Board of Agriculture, Sir John Sinclair introduced the word *statistics* into the English language in the 1798 publication of his book on a statistical account of Scotland. The word *statistics* is derived from the Latin word *status*, which is loosely defined as a statesman.

There are several reasons why you should study statistics.

1. Like professional people, you must be able to read and understand the various statistical studies performed in your fields. To have this understanding, you must be knowledgeable about the vocabulary, symbols, concepts, and statistical procedures used in these studies.
2. You may be called on to conduct research in your field, since statistical procedures are basic to research. To accomplish this, you must be able to design experiments; collect, organize, analyze, and summarize data; and possibly make reliable predictions or forecasts for future use. You must also be able to communicate the results of the study in your own words.
3. You can also use the knowledge gained from studying statistics to become better consumers and citizens. For example, you can make intelligent decisions about what products to purchase based on consumer studies, about government spending based on utilization studies, and so on.

It is the purpose of this chapter to introduce the goals for studying statistics by answering questions such as the following:

What are the branches of statistics?

What are data?

How are samples selected?

1–1 Descriptive and Inferential Statistics

OBJECTIVE 1

Demonstrate knowledge of statistical terms.

To gain knowledge about seemingly haphazard situations, statisticians collect information for *variables*, which describe the situation.

A variable is a characteristic or attribute that can assume different values.

Historical Note

The 1880 Census had so many questions on it that it took 10 years to publish the results.

Data are the values (measurements or observations) that the variables can assume. Variables whose values are determined by chance are called **random variables.**

Suppose that an insurance company studies its records over the past several years and determines that, on average, 3 out of every 100 automobiles the company insured were involved in accidents during a 1-year period. Although there is no way to predict the specific automobiles that will be involved in an accident (random occurrence), the company can adjust its rates accordingly, since the company knows the general pattern over the long run. (That is, on average, 3% of the insured automobiles will be involved in an accident each year.)

A collection of data values forms a **data set.** Each value in the data set is called a **data value** or a **datum.**

In statistics it is important to distinguish between a sample and a population.

A population consists of all subjects (human or otherwise) that are being studied.

Historical Note

The origin of descriptive statistics can be traced to data collection methods used in censuses taken by the Babylonians and Egyptians between 4500 and 3000 B.C. In addition, the Roman Emperor Augustus (27 B.C.–A.D. 17) conducted surveys on births and deaths of the citizens of the empire, as well as the number of livestock each owned and the crops each citizen harvested yearly.

When data are collected from every subject in the population, it is called a *census*.

For example, every 10 years the United States conducts a census. The primary purpose of this census is to determine the apportionment of the seats in the House of Representatives.

The first census was conducted in 1790 and was mandated by Article 1, Section 2 of the Constitution. As the United States grew, the scope of the census also grew. Today the Census limits questions to populations, housing, manufacturing, agriculture, and mortality. The Census is conducted by the Bureau of the Census, which is part of the Department of Commerce.

Most of the time, due to the expense, time, size of population, medical concerns, etc., it is not possible to use the entire population for a statistical study; therefore, researchers use samples.

A sample is a group of subjects selected from a population.

If the subjects of a sample are properly selected, most of the time they should possess the same or similar characteristics as the subjects in the population. See Figure 1–1.

However, the information obtained from a statistical sample is said to be *biased* if the results from the sample of a population are radically different from the results of a census of the population. Also, a sample is said to be biased if it does not represent the population from which it has been selected. The techniques used to properly select a sample are explained in Section 1–3.

OBJECTIVE 2

Differentiate between the two branches of statistics.

The body of knowledge called statistics is sometimes divided into two main areas, depending on how data are used. The two areas are

1. Descriptive statistics
2. Inferential statistics

FIGURE 1–1
Population and Sample

Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

In *descriptive statistics* the statistician tries to describe a situation. Consider the national census conducted by the U.S. government every 10 years. Results of this census give you the average age, income, and other characteristics of the U.S. population. To obtain this information, the Census Bureau must have some means to collect relevant data. Once data are collected, the bureau must organize and summarize them. Finally, the bureau needs a means of presenting the data in some meaningful form, such as charts, graphs, or tables.

The second area of statistics is called *inferential statistics*.

Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

Historical Note

Inferential statistics originated in the 1600s, when John Graunt published his book on population growth, *Natural and Political Observations Made upon the Bills of Mortality*. About the same time, another mathematician/astronomer, Edmond Halley, published the first complete mortality tables. (Insurance companies use mortality tables to determine life insurance rates.)

Here, the statistician tries to make inferences from *samples* to *populations*. Inferential statistics uses **probability**, i.e., the chance of an event occurring. You may be familiar with the concepts of probability through various forms of gambling. If you play cards, dice, bingo, or lotteries, you win or lose according to the laws of probability. Probability theory is also used in the insurance industry and other areas.

The area of inferential statistics called **hypothesis testing** is a decision-making process for evaluating claims about a population, based on information obtained from samples. For example, a researcher may wish to know if a new drug will reduce the number of heart attacks in men over age 70 years of age. For this study, two groups of men over age 70 would be selected. One group would be given the drug, and the other would be given a placebo (a substance with no medical benefits or harm). Later, the number of heart attacks occurring in each group of men would be counted, a statistical test would be run, and a decision would be made about the effectiveness of the drug.

Statisticians also use statistics to determine *relationships* among variables. For example, relationships were the focus of the most noted study in the 20th century, "Smoking and Health," published by the Surgeon General of the United States in 1964. He stated that after reviewing and evaluating the data, his group found a definite relationship between smoking and lung cancer. He did not say that cigarette smoking actually causes lung cancer, but that there is a relationship between smoking and lung cancer. This conclusion was based on a study done in 1958 by Hammond and Horn. In this study, 187,783 men were observed over a period of 45 months. The death rate from lung cancer in this group of volunteers was 10 times as great for smokers as for nonsmokers.

Unusual Stat

Twenty-nine percent of Americans want their boss's job.

Finally, by studying past and present data and conditions, statisticians try to make predictions based on this information. For example, a car dealer may look at past sales records for a specific month to decide what types of automobiles and how many of each type to order for that month next year.

EXAMPLE 1–1 Descriptive or Inferential Statistics

Determine whether descriptive or inferential statistics were used.

a. The average price of a 30-second ad for the Academy Awards show in a recent year was 1.90 million dollars.

b. The Department of Economic and Social Affairs predicts that the population of Mexico City, Mexico, in 2030 will be 238,647,000 people.

c. A medical report stated that taking statins is proven to lower heart attacks, but some people are at a slightly higher risk of developing diabetes when taking statins.

d. A survey of 2234 people conducted by the Harris Poll found that 55% of the respondents said that excessive complaining by adults was the most annoying social media habit.

SOLUTION

a. A descriptive statistic (average) was used since this statement was based on data obtained in a recent year.

b. Inferential statistics were used since this is a prediction for a future year.

c. Inferential statistics were used since this conclusion was drawn from data obtained from samples and used to conclude that the results apply to a population.

d. Descriptive statistics were used since this is a result obtained from a sample of 2234 survey respondents.

Applying the Concepts 1–1

Attendance and Grades

Read the following on attendance and grades, and answer the questions.

A study conducted at Manatee Community College revealed that students who attended class 95 to 100% of the time usually received an A in the class. Students who attended class 80 to 90% of the time usually received a B or C in the class. Students who attended class less than 80% of the time usually received a D or an F or eventually withdrew from the class.

Unusual Stat

Only one-third of crimes committed are reported to the police.

Based on this information, attendance and grades are related. The more you attend class, the more likely it is you will receive a higher grade. If you improve your attendance, your grades will probably improve. Many factors affect your grade in a course. One factor that you have considerable control over is attendance. You can increase your opportunities for learning by attending class more often.

1. What are the variables under study?
2. What are the data in the study?
3. Are descriptive, inferential, or both types of statistics used?
4. What is the population under study?
5. Was a sample collected? If so, from where?
6. From the information given, comment on the relationship between the variables.

See page 38 for the answers.

Exercises 1–1

1. Define statistics.
2. What is a variable?
3. What is meant by a census?
4. How does a population differ from a sample?
5. Explain the difference between descriptive and inferential statistics.
6. Name three areas where probability is used.
7. Why is information obtained from samples used more often than information obtained from populations?
8. What is meant by a biased sample?

For Exercises 9–17, determine whether descriptive or inferential statistics were used.

9. Because of the current economy, 49% of 18- to 34-year-olds have taken a job to pay the bills. (*Source:* Pew Research Center)
10. In 2025, the world population is predicted to be 8 billion people. (*Source:* United Nations)
11. In a weight loss study using teenagers at Boston University, 52% of the group said that they lost weight and kept it off by counting calories.
12. Based on a sample of 2739 respondents, it is estimated that pet owners spent a total of 14 billion dollars on veterinarian care for their pets. (*Source:* American Pet Products Association, Pet Owners Survey)
13. A recent article stated that over 38 million U.S. adults binge-drink alcohol.
14. The Centers for Disease Control and Prevention estimated that for a specific school year, 7% of children in kindergartens in the state of Oregon had nonmedical waivers for vaccinations.
15. A study conducted by a research network found that people with fewer than 12 years of education had lower life expectancies than those with more years of education.
16. A survey of 1507 smartphone users showed that 38% of them purchased insurance at the same time as they purchased their phones.
17. Forty-four percent of the people in the United States have type O blood. (*Source:* American Red Cross)

Extending the Concepts

18. Find three statistical studies and explain whether they used descriptive or inferential statistics.

19. Find a gambling game and explain how probability was used to determine the outcome.

1–2 Variables and Types of Data

OBJECTIVE 3

Identify types of data.

As stated in Section 1–1, statisticians gain information about a particular situation by collecting data for random variables. This section will explore in greater detail the nature of variables and types of data.

Variables can be classified as qualitative or quantitative.

Qualitative variables are variables that have distinct categories according to some

For example, if subjects are classified according to gender (male or female), then the variable *gender* is qualitative. Other examples of qualitative variables are religious preference and geographic locations.

Quantitative variables are variables that can be counted or measured.

For example, the variable *age* is numerical, and people can be ranked in order according to the value of their ages. Other examples of quantitative variables are heights, weights, and body temperatures.

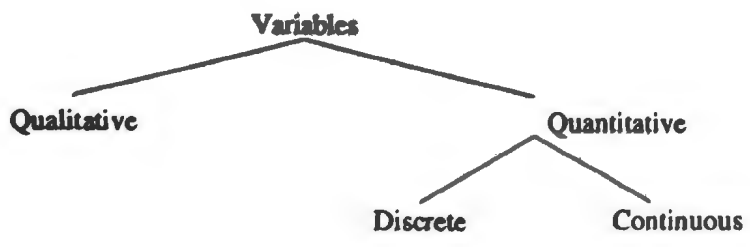
Quantitative variables can be further classified into two groups: discrete and continuous. *Discrete variables* can be assigned values such as 0, 1, 2, 3 and are said to be *countable*. Examples of discrete variables are the number of children in a family, the number of students in a classroom, and the number of calls received by a call center each day for a month.

Discrete variables assume values that can be counted.

Continuous variables, by comparison, can assume an infinite number of values in an interval between any two specific values. Temperature, for example, is a continuous variable, since the variable can assume an infinite number of values between any two given temperatures.

Continuous variables can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals.

The classification of variables can be summarized as follows:



EXAMPLE 1–2 Discrete or Continuous Data

Classify each variable as a discrete or continuous variable.

a. The number of hours during a week that children ages 12 to 15 reported that they watched television.

b. The number of touchdowns a quarterback scored each year in his college football career.

c. The amount of money a person earns per week working at a fast-food restaurant.

d. The weights of the football players on the teams that play in the NFL this year.

SOLUTION

a. Continuous, since the variable time is measured

b. Discrete, since the number of touchdowns is counted

c. Discrete, since the smallest value that money can assume is in cents

d. Continuous, since the variable weight is measured

Unusual Stat

Fifty-two percent of Americans live within 50 miles of a coastal shoreline.

Since continuous data must be measured, answers must be rounded because of the limits of the measuring device. Usually, answers are rounded to the nearest given unit. For example, heights might be rounded to the nearest inch, weights to the nearest ounce, etc. Hence, a recorded height of 73 inches could mean any measure from 72.5 inches up to but not including 73.5 inches. Thus, the boundary of this measure is given as 72.5–73.5 inches. The **boundary** of a number, then, is defined as a class in which a data value would be placed before the data value was rounded. *Boundaries are written for convenience as 72.5–73.5 but are understood to mean all values up to but not including 73.5.* Actual data values of 73.5 would be rounded to 74 and would be included in a class with boundaries of 73.5 up to but not including 74.5, written as 73.5–74.5. As another example, if a recorded weight is 86 pounds, the exact boundaries are 85.5 up to but not including 86.5, written as 85.5–86.5 pounds. Table 1–1 helps to clarify this concept. The boundaries of a continuous variable are given in one additional decimal place and always end with the digit 5.

TABLE 1–1 Recorded Values and Boundaries

Length	15 centimeters (cm)	14.5–15.5 cm
Temperature	86 degrees Fahrenheit (°F)	85.5–86.5°F
Time	0.43 second (sec)	0.425–0.435 sec
Mass	1.6 grams (g)	1.55–1.65 g

EXAMPLE 1–3 Class Boundaries

Find the boundaries for each measurement.

a. 17.6 inches

b. 23° Fahrenheit

c. 154.62 mg/dl

SOLUTION

a. 17.55–18.55 inches

b. 22.5–23.5° Fahrenheit

c. 154.615–154.625 mg/dl

In addition to being classified as qualitative or quantitative, variables can be classified by how they are categorized, counted, or measured. For example, can the data be organized into specific categories, such as area of residence (rural, suburban, or urban)? Can the data values be ranked, such as first place, second place, etc.? Or are the values obtained from measurement, such as heights, IQs, or temperature? This type of classification—i.e., how variables are categorized, counted, or measured—uses **measurement scales,** and four common types of scales are used: nominal, ordinal, interval, and ratio.

OBJECTIVE 4

Identify the measurement level for each variable.

The first level of measurement is called the *nominal level* of measurement. A sample of college instructors classified according to subject taught (e.g., English, history, psychology, or mathematics) is an example of nominal-level measurement. Classifying survey subjects as male or female is another example of nominal-level measurement. No ranking or order can be placed on the data. Classifying residents according to zip codes is also an example of the nominal level of measurement. Even though numbers are assigned as zip codes, there is no meaningful order or ranking. Other examples of nominal-level data are political party (Democratic, Republican, independent, etc.), religion (Christianity, Judaism, Islam, etc.), and marital status (single, married, divorced, widowed, separated).

Unusual Stat

Sixty-three percent of us say we would rather hear the bad news first.

The nominal level of measurement classifies data into mutually exclusive (nonoverlapping) categories in which no order or ranking can be imposed on the data.

Historical Note

When data were first analyzed statistically by Karl Pearson and Francis Galton, almost all were continuous data. In 1899, Pearson began to analyze discrete data. Pearson found that some data, such as eye color, could not be measured, so he termed such data *nominal data.* Ordinal data were introduced by a German numerologist Frederich Mohs in 1822 when he introduced a hardness scale for minerals. For example, the hardest stone is the diamond, which he assigned a hardness value of 1500. Quartz was assigned a hardness value of 100. This does not mean that a diamond is 15 times harder than quartz. It only means that a diamond is harder than quartz. In 1947, a psychologist named Stanley Smith Stevens made a further division of continuous data into two categories, namely, interval and ratio.

The next level of measurement is called the *ordinal level.* Data measured at this level can be placed into categories, and these categories can be ordered, or ranked. For example, from student evaluations, guest speakers might be ranked as superior, average, or poor. Floats in a homecoming parade might be ranked as first place, second place, etc. *Note that precise measurement of differences in the ordinal level of measurement* does not *exist.* For instance, when people are classified according to their build (small, medium, or large), a large variation exists among the individuals in each class.

Other examples of ordinal data are letter grades (A, B, C, D, F).

The ordinal level of measurement classifies data into categories that can be ranked; however, precise differences between the ranks do not exist.

The third level of measurement is called the *interval level.* This level differs from the ordinal level in that precise differences do exist between units. For example, many standardized psychological tests yield values measured on an interval scale. IQ is an example of such a variable. There is a meaningful difference of 1 point between an IQ of 109 and an IQ of 110. Temperature is another example of interval measurement, since there is a meaningful difference of 1°F between each unit, such as 72 and 73°F. *One property is lacking in the interval scale: There is no true zero.* For example, IQ tests do not measure people who have no intelligence. For temperature, 0°F does not mean no heat at all.

The interval level of measurement ranks data, and precise differences between units of measure do exist; however, there is no meaningful zero.

The final level of measurement is called the *ratio level.* Examples of ratio scales are those used to measure height, weight, area, and number of phone calls received. Ratio scales have differences between units (1 inch, 1 pound, etc.) and a true zero. In addition, the ratio scale contains a true ratio between values. For example, if one person can lift 200 pounds and another can lift 100 pounds, then the ratio between them is 2 to 1. Put another way, the first person can lift twice as much as the second person.

The ratio level of measurement possesses all the characteristics of interval measurement, and there exists a true zero. In addition, true ratios exist when the same variable is measured on two different members of the population.

TABLE 1–2 Examples of Measurement Scales

Nominal-level data	Ordinal-level data	Interval-level data	Ratio-level data
Zip code Gender (male, female) Eye color (blue, brown, green, hazel) Political affiliation Religious affiliation Major field (mathematics, computers, etc.) Nationality	Grade (A, B, C, D, F) Judging (first place, second place, etc.) Rating scale (poor, good, excellent) Ranking of tennis players	SAT score IQ Temperature	Height Weight Time Salary Age

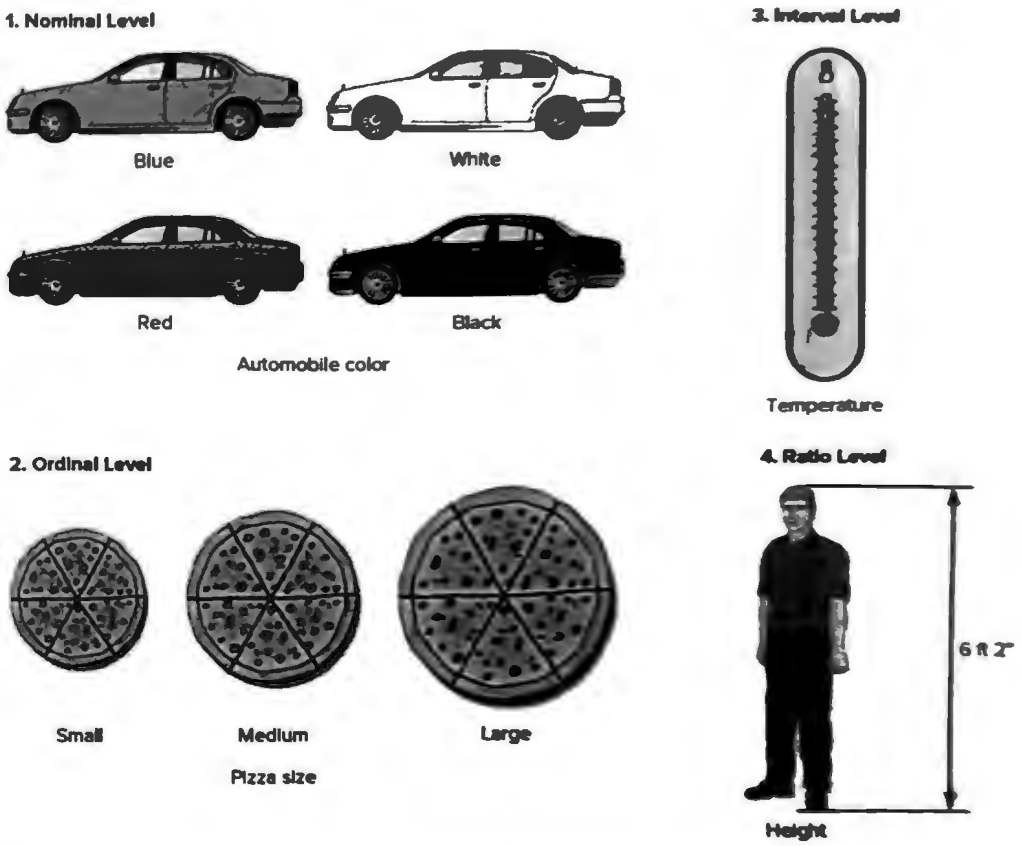


FIGURE 1–2
Measurement Scales

There is not complete agreement among statisticians about the classification of data into one of the four categories. For example, some researchers classify IQ data as ratio data rather than interval. Also, data can be altered so that they fit into a different category. For instance, if the incomes of all professors of a college are classified into the three categories of low, average, and high, then a ratio variable becomes an ordinal variable. Table 1–2 gives some examples of each type of data. See Figure 1–2.

EXAMPLE 1–4 Measurement Levels

What level of measurement would be used to measure each variable?

a. The ages of authors who wrote the hardback versions of the top 25 fiction books sold during a specific week

b. The colors of baseball hats sold in a store for a specific year

c. The highest temperature for each day of a specific month

d. The ratings of bands that played in the homecoming parade at a college

SOLUTION

a. Ratio

b. Nominal

c. Interval

d. Ordinal

Applying the Concepts 1–2

Fatal Transportation Injuries

Read the following information about the number of fatal accidents for the transportation industry in for a specific year, and answer each question.

Highway accidents	968
Railway accidents	44
Water vehicle accidents	52
Aircraft accidents	151

Source: Bureau of Labor Statistics.

1. Name the variables under study.
2. Categorize each variable as quantitative or qualitative.
3. Categorize each quantitative variable as discrete or continuous.
4. Identify the level of measurement for each variable.
5. The railroad had the fewest fatalities for the specific year. Does that mean railroads have fewer accidents than the other industries?
6. What factors other than safety influence a person's choice of transportation?
7. From the information given, comment on the relationship between the variables.

See page 38 for the answers.

Exercises 1–2

1. Explain the difference between qualitative variables and quantitative variables.
2. Explain the difference between discrete and continuous variables.
3. Why are continuous variables rounded when they are used in statistical studies?
4. Name and define the four types of measurement levels used in statistics.

For Exercises 5–10, determine whether the data are qualitative or quantitative.

5. Sizes of soft drinks sold by a fast-food restaurant (small, medium, and large)
6. Pizza sizes (small, medium, and large)
7. Cholesterol counts for individuals
8. Microwave wattage
9. Number of degrees awarded by a college each year for the last 10 years
10. Ratings of teachers

For Exercises 11–16, determine whether the data are discrete or continuous.

11. Number of phone calls received by a 911 call center each day
12. Systolic blood pressure readings
13. Weights of the suitcases of airline passengers on a specific flight
14. Votes received by mayoral candidates in a city election
15. Number of students in the mathematics classes during the fall semester at your school for a particular school year
16. Temperatures at a seashore resort

Chapter Quiz

Determine whether each statement is true or false. If the statement is false, explain why.

1. Probability is used as a basis for inferential statistics.
2. When the sample does not represent the population, it is called a biased sample.
3. The difference between a sampling measure and a population measure is called a nonsampling error.
4. When the population of college professors is divided into groups according to their rank (instructor, assistant professor, etc.) and then several are selected from each group to make up a sample, the sample is called a cluster sample.
5. The variable temperature is an example of a quantitative variable.
6. The height of basketball players is considered a continuous variable.
7. The boundary of a value such as 6 inches would be 5.9–6.1 inches.

Select the best answer.

8. The number of ads on a one-hour television show is what type of data?
   a. Nominal
   b. Qualitative
   c. Discrete
   d. Continuous
9. What are the boundaries of 25.6 ounces?
   a. 25–26 ounces
   b. 25.55–25.65 ounces
   c. 25.5–25.7 ounces
   d. 20–39 ounces
10. A researcher divided subjects into two groups according to gender and then selected members from each group for her sample. What sampling method was the researcher using?
    a. Cluster
    b. Random
    c. Systematic
    d. Stratified
11. Data that can be classified according to color are measured on what scale?
    a. Nominal
    b. Ratio
    c. Ordinal
    d. Interval
12. A study that involves no researcher intervention is called
    a. An experimental study.
    b. A noninvolvement study.
    c. An observational study.
    d. A quasi-experimental study.
13. A variable that interferes with other variables in the study is called
    a. A confounding variable.
    b. An explanatory variable.
    c. An outcome variable.
    d. An interfering variable.

Use the best answer to complete these statements.

14. Two major branches of statistics are _____ and _____.
15. Two uses of probability are _____ and _____
16. The group of all subjects under study is called a(n) _____.
17. A group of subjects selected from the group of all subjects under study is called a(n) _____.
18. Three reasons why samples are used in statistics:
    a. _____ *b.* _____ *c.* _____
19. The four basic sampling methods are
    a. _____ *b.* _____ *c.* _____ *d.* _____
20. A study that uses intact groups when it is not possible to randomly assign participants to the groups is called a(n) _____ study.
21. In a research study, participants should be assigned to groups using _____ methods, if possible.
22. For each statement, decide whether descriptive or inferential statistics is used.
    a. The average life expectancy in New Zealand is 78.49 years (*Source: World Factbook*).
    b. A diet high in fruits and vegetables will lower blood pressure (*Source:* Institute of Medicine).
    c. The total amount of estimated losses for Hurricane Katrina was $125 billion (*Source: The World Almanac and Book of Facts*).
    d. Researchers stated that the shape of a person's ears is relative to the person's aggression (*Source: American Journal of Human Biology*).

e. In 2050, it is estimated that there will be 18 million Americans who are age 85 and over (*Source:* U.S. Census Bureau).

23. Classify each as nominal-level, ordinal-level, interval-level, or ratio-level of measurement.

a. Rating of movies as G, PG, and R
b. Number of candy bars sold on a fund drive
c. Classification of automobiles as subcompact, compact, standard, and luxury
d. Temperatures of hair dryers
e. Weights of suitcases on a commercial airliner

24. Classify each variable as discrete or continuous.

a. Ages of people working in a large factory
b. Number of cups of coffee served at a restaurant
c. The amount of drug injected into a guinea pig
d. The time it takes a student to drive to school
e. The number of gallons of milk sold each day at a grocery store

25. Give the boundaries of each.

a. 32 minutes
b. 0.48 millimeter
c. 6.2 inches
d. 19 pounds
e. 12.1 quarts

Critical Thinking Challenges

1. World's Busiest Airports A study of the world's busiest airports was conducted by *Airports Council International*. Describe three variables that one could use to determine which airports are the busiest. What *units* would one use to measure these variables? Are these variables categorical, discrete, or continuous?

2. Smoking and Criminal Behavior The results of a study published in *Archives of General Psychiatry* stated that male children born to women who smoke during pregnancy run a risk of violent and criminal behavior that lasts into adulthood. The results of this study were challenged by some people in the media. Give several reasons why the results of this study would be challenged.

3. Piano Lessons Improve Math Ability The results of a study published in *Neurological Research* stated that second-graders who took piano lessons and played a computer math game more readily grasped math problems in fractions and proportions than a similar group who took an English class and played the same math game. What type of inferential study was this? Give several reasons why the piano lessons could improve a student's math ability.

4. ACL Tears in Collegiate Soccer Players A study of 2958 collegiate soccer players showed that in 46 anterior cruciate ligament (ACL) tears, 36 were in women. Calculate the percentages of tears for each gender.

a. Can it be concluded that female athletes tear their knees more often than male athletes?
b. Comment on how this study's conclusion might have been reached.

Data Projects

1. Business and Finance Investigate the types of data that are collected regarding stock and bonds, for example, price, earnings ratios, and bond ratings. Find as many types of data as possible. For each, identify the level of measurement as nominal, ordinal, interval, or ratio. For any quantitative data, also note if they are discrete or continuous.

2. Sports and Leisure Select a professional sport. Investigate the types of data that are collected about that sport, for example, in baseball, the level of play (A, AA, AAA, Major League), batting average, and home-runs. For each, identify the level of measurement as nominal, ordinal, interval, or ratio. For any quantitative data, also note if they are discrete or continuous.

3. Technology Music organization programs on computers and music players maintain information about a song, such as the writer, song length, genre, and your personal rating. Investigate the types of data collected about a song. For each, identify the level of measurement as nominal, ordinal, interval, or ratio. For any quantitative data, also note if they are discrete or continuous.

2

Frequency Distributions and Graphs

© Image Source, all rights reserved. RF

STATISTICS TODAY

How Your Identity Can Be Stolen

Identity fraud is a big business today—more than 12.7 million people were victims. The total amount of the fraud in 2014 was $16 billion. The average amount of the fraud for a victim is $1260, and the average time to correct the problem is 40 hours. The ways in which a person's identity can be stolen are presented in the following table:

Government documents or benefits fraud	38.7%
Credit card fraud	17.4
Phone or utilities fraud	12.5
Bank fraud	8.2
Attempted identity theft	4.8
Employment-related fraud	4.8
Loan fraud	4.4
Other identity theft	9.2

Source: Javelin Strategy & Research; Council of Better Business Bureau, Inc.

Looking at the numbers presented in a table does not have the same impact as presenting numbers in a well-drawn chart or graph. The article did not include any graphs. This chapter will show you how to construct appropriate graphs to represent data and help you to get your point across to your audience.

See Statistics Today—Revisited at the end of the chapter for some suggestions on how to represent the data graphically.

OUTLINE

Introduction

2–1 Organizing Data

2–2 Histograms, Frequency Polygons, and Ogives

2–3 Other Types of Graphs

Summary

OBJECTIVES

After completing this chapter, you should be able to

1. Organize data using a frequency distribution.
2. Represent data in frequency distributions graphically, using histograms, frequency polygons, and ogives.
3. Represent data using bar graphs, Pareto charts, time series graphs, pie graphs, and dotplots.
4. Draw and interpret a stem and leaf plot.

Introduction

When conducting a statistical study, the researcher must gather data for the particular variable under study. For example, if a researcher wishes to study the number of people who were bitten by poisonous snakes in a specific geographic area over the past several years, he or she has to gather the data from various doctors, hospitals, or health departments.

To describe situations, draw conclusions, or make inferences about events, the researcher must organize the data in some meaningful way. The most convenient method of organizing data is to construct a *frequency distribution.*

After organizing the data, the researcher must present them so they can be understood by those who will benefit from reading the study. The most useful method of presenting the data is by constructing *statistical charts* and *graphs.* There are many different types of charts and graphs, and each one has a specific purpose.

This chapter explains how to organize data by constructing frequency distributions and how to present the data by constructing charts and graphs. The charts and graphs illustrated here are histograms, frequency polygons, ogives, pie graphs, Pareto charts, and time series graphs. A graph that combines the characteristics of a frequency distribution and a histogram, called a stem and leaf plot, is also explained.

2–1 Organizing Data

OBJECTIVE

Organize data using a frequency distribution.

Suppose a researcher wished to do a study on the ages of the 50 wealthiest people in the world. The researcher first would have to get the data on the ages of the people. In this case, these ages are listed in *Forbes Magazine.* When the data are in original form, they are called **raw data** and are listed next.

45	46	64	57	85
92	51	71	54	48
27	66	76	55	69
54	44	54	75	46
61	68	78	61	83
88	45	89	67	56
81	58	55	62	38
55	56	64	81	38
49	68	91	56	68
46	47	83	71	62

Since little information can be obtained from looking at raw data, the researcher organizes the data into what is called a *frequency distribution.*

Unusual Stats

Of Americans 50 years old and over, 23% think their greatest achievements are still ahead of them.

[redacted]

Each raw data value is placed into a quantitative or qualitative category called a **class.** The **frequency** of a class then is the number of data values contained in a specific class. A frequency distribution is shown for the preceding data set.

27–35	/	1
36–44	///	3
45–53	𝍸 ////	9
54–62	𝍸 𝍸 𝍸	15
63–71	𝍸 𝍸	10
72–80	///	3
81–89	𝍸 //	7
90–98	//	2
		50

Now some general observations can be made from looking at the frequency distribution. For example, it can be stated that the majority of the wealthy people in the study are 45 years old or older.

The classes in this distribution are 27–35, 36–44, etc. These values are called *class limits*. The data values 27, 28, 29, 30, 31, 32, 33, 34, 35 can be tallied in the first class; 36, 37, 38, 39, 40, 41, 42, 43, 44 in the second class; and so on.

Two types of frequency distributions that are most often used are the *categorical frequency distribution* and the *grouped frequency distribution*. The procedures for constructing these distributions are shown now.

Categorical Frequency Distributions

The categorical frequency distribution is used for data that can be placed in specific categories, such as nominal- or ordinal-level data. For example, data such as political affiliation, religious affiliation, or major field of study would use categorical frequency distributions.

EXAMPLE 2–1 Distribution of Blood Types

Twenty-five army inductees were given a blood test to determine their blood type. The data set is

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construct a frequency distribution for the data.

SOLUTION

Since the data are categorical, discrete classes can be used. There are four blood types: A, B, O, and AB. These types will be used as the classes for the distribution.

The procedure for constructing a frequency distribution for categorical data is given next.

Step 1 Make a table as shown.

A			
B			
O			
AB			

Step 2 Tally the data and place the results in column B.

Step 3 Count the tallies and place the results in column C.

Step 4 Find the percentage of values in each class by using the formula

$$\% = \frac{f}{n} \cdot 100$$

where f = frequency of the class and n = total number of values. For example, in the class of type A blood, the percentage is

$$\% = \frac{5}{25} \cdot 100 = 20\%$$

Percentages are not normally part of a frequency distribution, but they can be added since they are used in certain types of graphs such as pie graphs. Also, the decimal equivalent of a percent is called a *relative frequency*.

Step 5 Find the totals for columns C (frequency) and D (percent). The completed table is shown. It is a good idea to add the percent column to make sure it sums to 100%. This column won't always sum to 100% because of rounding.

A	𝍸	5	20
B	𝍸 //	7	28
O	𝍸 ////	9	36
AB	////	4	16
		Total 25	100%

For the sample, more people have type O blood than any other type.

Grouped Frequency Distributions

When the range of the data is large, the data must be grouped into classes that are more than one unit in width, in what is called a **grouped frequency distribution.** For example, a distribution of the blood glucose levels in milligrams per deciliter (mg/dL) for 50 randomly selected college students is shown.

Unusual Stats

Six percent of Americans say they find life dull.

58–64	57.5–64.5	/	1
65–71	64.5–71.5	𝍸 /	6
72–78	71.5–78.5	𝍸 𝍸	10
79–85	78.5–85.5	𝍸 𝍸 ////	14
86–92	85.5–92.5	𝍸 𝍸 //	12
93–99	92.5–99.5	𝍸	5
100–106	99.5–106.5	//	2
			Total 50

The procedure for constructing the preceding frequency distribution is given in Example 2–2; however, several things should be noted. In this distribution, the values 58 and 64 of the first class are called *class limits*. The **lower class limit** is 58; it represents the smallest data value that can be included in the class. The **upper class limit** is 64; it

represents the largest data value that can be included in the class. The numbers in the second column are called **class boundaries.** These numbers are used to separate the classes so that there are no gaps in the frequency distribution. The gaps are due to the limits; for example, there is a gap between 64 and 65.

Students sometimes have difficulty finding class boundaries when given the class limits. The basic rule of thumb is that *the class limits should have the same decimal place value as the data, but the class boundaries should have one additional place value and end in a 5.* For example, if the values in the data set are whole numbers, such as 59, 68, and 82, the limits for a class might be 58–64, and the boundaries are 57.5–64.5. Find the boundaries by subtracting 0.5 from 58 (the lower class limit) and adding 0.5 to 64 (the upper class limit).

$$\text{Lower limit} - 0.5 = 58 - 0.5 = 57.5 = \text{lower boundary}$$
$$\text{Upper limit} + 0.5 = 64 + 0.5 = 64.5 = \text{upper boundary}$$

Unusual Stats

One out of every hundred people in the United States is color-blind.

If the data are in tenths, such as 6.2, 7.8, and 12.6, the limits for a class hypothetically might be 7.8–8.8, and the boundaries for that class would be 7.75–8.85. Find these values by subtracting 0.05 from 7.8 and adding 0.05 to 8.8.

Class boundaries are not always included in frequency distributions; however, they give a more formal approach to the procedure of organizing data, including the fact that sometimes the data have been rounded. You should be familiar with boundaries since you may encounter them in a statistical study.

Finally, the **class width** for a class in a frequency distribution is found by subtracting the lower (or upper) class limit of one class from the lower (or upper) class limit of the next class. For example, the class width in the preceding distribution on the distribution of blood glucose levels is 7, found from $65 - 58 = 7$.

The class width can also be found by subtracting the lower boundary from the upper boundary for any given class. In this case, $64.5 - 57.5 = 7$.

Note: Do not subtract the limits of a single class. It will result in an incorrect answer.

The researcher must decide how many classes to use and the width of each class. To construct a frequency distribution, follow these rules:

1. *There should be between 5 and 20 classes.* Although there is no hard-and-fast rule for the number of classes contained in a frequency distribution, it is of utmost importance to have enough classes to present a clear description of the collected data.
2. *It is preferable but not absolutely necessary that the class width be an odd number.* This ensures that the midpoint of each class has the same place value as the data. The **class midpoint** X_m is obtained by adding the lower and upper boundaries and dividing by 2, or adding the lower and upper limits and dividing by 2:

$$X_m = \frac{\text{lower boundary} + \text{upper boundary}}{2}$$

or

$$X_m = \frac{\text{lower limit} + \text{upper limit}}{2}$$

For example, the midpoint of the first class in the example with glucose levels is

$$\frac{57.5 + 64.5}{2} = 61 \quad \text{or} \quad \frac{58 + 64}{2} = 61$$

The midpoint is the numeric location of the center of the class. Midpoints are necessary for graphing (see Section 2–2). If the class width is an even number, the

midpoint is in tenths. For example, if the class width is 6 and the boundaries are 5.5 and 11.5, the midpoint is

$$\frac{5.5 + 11.5}{2} = \frac{17}{2} = 8.5$$

Rule 2 is only a suggestion, and it is not rigorously followed, especially when a computer is used to group data.

3. *The classes must be mutually exclusive.* Mutually exclusive classes have nonoverlapping class limits so that data cannot be placed into two classes. Many times, frequency distributions such as this

Age
10–20
20–30
30–40
40–50

are found in the literature or in surveys. If a person is 40 years old, into which class should she or he be placed? A better way to construct a frequency distribution is to use classes such as

Age
10–20
21–31
32–42
43–53

Recall that boundaries are mutually exclusive. For example, when a class boundary is 5.5 to 10.5, the data values that are included in that class are values from 6 to 10. A data value of 5 goes into the previous class, and a data value of 11 goes into the next-higher class.

4. *The classes must be continuous.* Even if there are no values in a class, the class must be included in the frequency distribution. There should be no gaps in a frequency distribution. The only exception occurs when the class with a zero frequency is the first or last class. A class with a zero frequency at either end can be omitted without affecting the distribution.

5. *The classes must be exhaustive.* There should be enough classes to accommodate all the data.

6. *The classes must be equal in width.* This avoids a distorted view of the data.

One exception occurs when a distribution has a class that is **open-ended.** That is, the first class has no specific lower limit, or the last class has no specific upper limit. A frequency distribution with an open-ended class is called an **open-ended distribution.** Here are two examples of distributions with open-ended classes.

10–20	3
21–31	6
32–42	4
43–53	10
54 and above	8

Below 110	16
110–114	24
115–119	38
120–124	14
125–129	5

The frequency distribution for age is open-ended for the last class, which means that anybody who is 54 years or older will be tallied in the last class. The distribution for minutes is open-ended for the first class, meaning that any minute values below 110 will be tallied in that class.

The steps for constructing a grouped frequency distribution are summarized in the following Procedure Table.

Procedure Table

Constructing a Grouped Frequency Distribution

Step 1 Determine the classes.
- Find the highest and lowest values.
- Find the range.
- Select the number of classes desired.
- Find the width by dividing the range by the number of classes and rounding up.
- Select a starting point (usually the lowest value or any convenient number less than the lowest value); add the width to get the lower limits.
- Find the upper class limits.
- Find the boundaries.

Step 2 Tally the data.

Step 3 Find the numerical frequencies from the tallies, and find the cumulative frequencies.

Example 2–2 shows the procedure for constructing a grouped frequency distribution, i.e., when the classes contain more than one data value.

EXAMPLE 2–2 Record High Temperatures

These data represent the record high temperatures in degrees Fahrenheit (°F) for each of the 50 states. Construct a grouped frequency distribution for the data, using 7 classes.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Source: The World Almanac and Book of Facts.

SOLUTION

The procedure for constructing a grouped frequency distribution for numerical data follows.

Unusual Stats

America's most popular beverages are soft drinks. It is estimated that, on average, each person drinks about 52 gallons of soft drinks per year, compared to 22 gallons of beer.

Step 1 Determine the classes.

Find the highest value and lowest value: $H = 134$ and $L = 100$.

Find the range: R = highest value − lowest value $= H - L$, so

$$R = 134 - 100 = 34$$

Select the number of classes desired (usually between 5 and 20). In this case, 7 is arbitrarily chosen.

Find the class width by dividing the range by the number of classes.

$$\text{Width} = \frac{R}{\text{number of classes}} = \frac{34}{7} = 4.9$$

Round the answer up to the nearest whole number if there is a remainder: $4.9 \approx 5$. (Rounding *up* is different from rounding *off*. A number is rounded up if there is any decimal remainder when dividing. For example, $85 \div 6 = 14.167$ and is rounded up to 15. Also, $53 \div 4 = 13.25$ and is rounded up to 14. (Also, after dividing, if there is no remainder, you will need to add an extra class to accommodate all the data.)

Select a starting point for the lowest class limit. This can be the smallest data value or any convenient number less than the smallest data value. In this case, 100 is used. Add the width to the lowest score taken as the starting point to get the lower limit of the next class. Keep adding until there are 7 classes, as shown, 100, 105, 110, etc.

Historical Note

Florence Nightingale, a nurse in the Crimean War in 1854, used statistics to persuade government officials to improve hospital care of soldiers in order to reduce the death rate from unsanitary conditions in the military hospitals that cared for the wounded soldiers.

Subtract one unit from the lower limit of the second class to get the upper limit of the first class. Then add the width to each upper limit to get all the upper limits.

$$105 - 1 = 104$$

The first class is 100–104, the second class is 105–109, etc.

Find the class boundaries by subtracting 0.5 from each lower class limit and adding 0.5 to each upper class limit:

99.5–104.5, 104.5–109.5, etc.

Step 2 Tally the data.

Step 3 Find the numerical frequencies from the tallies.

The completed frequency distribution is

100–104	99.5–104.5	//	2
105–109	104.5–109.5	𝍸 ///	8
110–114	109.5–114.5	𝍸 𝍸 𝍸 ///	18
115–119	114.5–119.5	𝍸 𝍸 ///	13
120–124	119.5–124.5	𝍸 //	7
125–129	124.5–129.5	/	1
130–134	129.5–134.5	/	1
			Total 50

The frequency distribution shows that the class 109.5–114.5 contains the largest number of temperatures (18) followed by the class 114.5–119.5 with 13 temperatures. Hence, most of the temperatures (31) fall between 110 and 119°F.

Sometimes it is necessary to use a *cumulative frequency distribution*. A **cumulative frequency distribution** is a distribution that shows the number of data values less than or equal to a specific value (usually an upper boundary). The values are found by adding the frequencies of the classes less than or equal to the upper class boundary of a specific class. This gives an ascending cumulative frequency. In this example, the cumulative frequency for the first class is $0 + 2 = 2$; for the second class it is $0 + 2 + 8 = 10$; for the third class it is $0 + 2 + 8 + 18 = 28$. Naturally, a shorter way to do this would be to just add the cumulative frequency of the class below to the frequency of the given class. For example, the cumulative frequency for the number of data values less than 114.5 can be

found by adding 10 + 18 = 28. The cumulative frequency distribution for the data in this example is as follows:

Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Cumulative frequencies are used to show how many data values are accumulated up to and including a specific class. In Example 2–2, of the total record high temperatures 28 are less than or equal to 114°F. Forty-eight of the total record high temperatures are less than or equal to 124°F.

After the raw data have been organized into a frequency distribution, it will be analyzed by looking for peaks and extreme values. The peaks show which class or classes have the most data values compared to the other classes. Extreme values, called *outliers*, show large or small data values that are relative to other data values.

When the range of the data values is relatively small, a frequency distribution can be constructed using single data values for each class. This type of distribution is called an **ungrouped frequency distribution** and is shown next.

EXAMPLE 2–3 Hours of Sleep

The data shown represent the number of hours 30 college students said they sleep per night. Construct and analyze a frequency distribution.

8	6	6	8	5	7
7	8	7	6	6	7
9	7	7	6	8	10
6	7	6	7	8	7
7	8	7	8	9	8

SOLUTION

Step 1 Determine the number of classes. Since the range is small (10 − 5 = 5), classes consisting of a single data value can be used. They are 5, 6, 7, 8, 9, and 10.

Note: If the data are continuous, class boundaries can be used. Subtract 0.5 from each class value to get the lower class boundary, and add 0.5 to each class value to get the upper class boundary.

Step 2 Tally the data.

Step 3 From the tallies, find the numerical frequencies and cumulative frequencies. The completed ungrouped frequency distribution is shown.

Class limits	Class boundaries	Tally	Frequency
5	4.5–5.5	/	1
6	5.5–6.5	𝍸 //	7
7	6.5–7.5	𝍸 𝍸 /	11
8	7.5–8.5	𝍸 ///	8
9	8.5–9.5	//	2
10	9.5–10.5	/	1

In this case, 11 students sleep 7 hours a night. Most of the students sleep between 5.5 and 8.5 hours.

The cumulative frequencies are

	Cumulative frequency
Less than 4.5	0
Less than 5.5	1
Less than 6.5	8
Less than 7.5	19
Less than 8.5	27
Less than 9.5	29
Less than 10.5	30

When you are constructing a frequency distribution, the guidelines presented in this section should be followed. However, you can construct several different but correct frequency distributions for the same data by using a different class width, a different number of classes, or a different starting point.

Interesting Fact

Male dogs bite children more often than female dogs do, however, female cats bite children more often than male cats do.

Furthermore, the method shown here for constructing a frequency distribution is not unique, and there are other ways of constructing one. Slight variations exist, especially in computer packages. But regardless of what methods are used, classes should be mutually exclusive, continuous, exhaustive, and of equal width.

In summary, the different types of frequency distributions were shown in this section. The first type, shown in Example 2–1, is used when the data are categorical (nominal), such as blood type or political affiliation. This type is called a categorical frequency distribution. The second type of distribution is used when the range is large and classes several units in width are needed. This type is called a grouped frequency distribution and is shown in Example 2–2. Another type of distribution is used for numerical data and when the range of data is small, as shown in Example 2–3. Since each class is only one unit, this distribution is called an ungrouped frequency distribution.

All the different types of distributions are used in statistics and are helpful when one is organizing and presenting data.

The reasons for constructing a frequency distribution are as follows:

1. To organize the data in a meaningful, intelligible way.
2. To enable the reader to determine the nature or shape of the distribution.
3. To facilitate computational procedures for measures of average and spread (shown in Sections 3–1 and 3–2).

4. To enable the researcher to draw charts and graphs for the presentation of data (shown in Section 2–2).
5. To enable the reader to make comparisons among different data sets.

The factors used to analyze a frequency distribution are essentially the same as those used to analyze histograms and frequency polygons, which are shown in Section 2–2.

Applying the Concepts 2–1

Ages of Presidents at Inauguration

The data represent the ages of our Presidents at the time they were first inaugurated.

57	61	57	57	58	57	61	54	68
51	49	64	50	48	65	52	56	46
54	49	51	47	55	55	54	42	51
56	55	51	54	51	60	62	43	55
56	61	52	69	64	46	54	47	

1. Were the data obtained from a population or a sample? Explain your answer.
2. What was the age of the oldest President?
3. What was the age of the youngest President?
4. Construct a frequency distribution for the data. (Use your own judgment as to the number of classes and class size.)
5. Are there any peaks in the distribution?
6. Identify any possible outliers.
7. Write a brief summary of the nature of the data as shown in the frequency distribution.

See page 108 for the answers.

Exercises 2–1

1. List five reasons for organizing data into a frequency distribution.
2. Name the three types of frequency distributions, and explain when each should be used.
3. How many classes should frequency distributions have? Why should the class width be an odd number?
4. What are open-ended frequency distributions? Why are they necessary?

For Exercises 5–8, find the class boundaries, midpoints, and widths for each class.

5. 58–62
6. 125–131
7. 16.35–18.46
8. 16.3–18.5

For Exercises 9–12, show frequency distributions that are incorrectly constructed. State the reasons why they are wrong.

9.

Class	Frequency
10–19	1
20–29	2
30–34	0
35–45	5
46–51	8

10.

Class	Frequency
5–9	1
9–13	2
13–17	5
17–20	6
20–24	3

11.

Class	Frequency
162–164	3
165–167	7
168–170	18
174–176	0
177–179	5

12.

Class	Frequency
9–13	1
14–19	6
20–25	2
26–28	5
29–32	9

13. Favorite Coffee Flavor A survey was taken asking the favorite flavor of a coffee drink a person prefers. The responses were V = Vanilla, C = Caramel, M = Mocha, H = Hazelnut, and P = Plain. Construct a categorical frequency distribution for the data. Which class has the most data values and which class has the fewest data values?

V	C	P	P	M	M	P	P	M	C
M	M	V	M	M	M	V	M	M	M
P	V	C	M	V	M	C	P	M	P
M	M	M	P	M	M	C	V	M	C
C	P	M	P	M	H	H	P	H	P

14. Trust in Internet Information A survey was taken on how much trust people place in the information they read on the Internet. Construct a categorical frequency distribution for the data. A = trust in all that they read, M = trust in most of what they read, H = trust in about one-half of what they read, S = trust in a small portion of what they read. (Based on information from the *UCLA Internet Report.*)

M	M	M	A	H	M	S	M	H	M
S	M	M	M	M	A	M	M	A	M
M	M	H	M	M	M	H	M	H	M
A	M	M	M	H	M	M	M	M	M

15. Eating at Fast Food Restaurants A survey was taken of 50 individuals. They were asked how many days per week they ate at a fast-food restaurant. Construct a frequency distribution using 8 classes (0–7). Based on the distribution, how often did most people eat at a fast-food restaurant?

1	3	4	0	4
5	2	2	3	1
2	2	2	2	2
2	2	2	2	3
2	2	5	2	4
2	4	5	2	1
4	1	3	2	2
2	0	7	2	3
2	2	2	5	2
3	3	4	1	3

16. Ages of Dogs The ages of 20 dogs in a pet shelter are shown. Construct a frequency distribution using 7 classes.

5	8	7	6	3
9	4	4	5	8
7	4	7	5	7
3	5	8	4	9

17. Maximum Wind Speeds The data show the maximum wind speeds in miles per hour recorded for 40 states. Construct a frequency distribution using 7 classes.

59	78	62	72	67
76	92	77	64	83
64	70	67	75	75
78	75	71	72	93
68	69	76	72	85
64	70	77	74	72
53	67	48	76	59
87	53	77	70	63

Source: NOAA

18. Stories in the World's Tallest Buildings The number of stories in each of a sample of the world's 30 tallest buildings follows. Construct a grouped frequency distribution and a cumulative frequency distribution with 7 classes.

88	88	110	88	80	69	102	78	70	55
79	85	80	100	60	90	77	55	75	55
54	60	75	64	105	56	71	70	65	72

Source: New York Times Almanac.

19. Ages of Declaration of Independence Signers The ages of the signers of the Declaration of Independence are shown. (Age is approximate since only the birth year appeared in the source, and one has been omitted since his birth year is unknown.) Construct a grouped frequency distribution and a cumulative frequency distribution for the data, using 7 classes.

41	54	47	40	39	35	50	37	49	42	70	32
44	52	39	50	40	30	34	69	39	45	33	42
44	63	60	27	42	34	50	42	52	38	36	45
35	43	48	46	31	27	55	63	46	33	60	62
35	46	45	34	53	50	50					

Source: The Universal Almanac.

20. Salaries of Governors Here are the salaries (in dollars) of the governors of 25 randomly selected states. Construct a grouped frequency distribution with 6 classes.

112,895	117,312	140,533	110,000	115,331
95,000	177,500	120,303	139,590	150,000
173,987	130,000	133,821	144,269	142,542
150,000	145,885	105,000	93,600	166,891
130,273	70,000	113,834	117,817	137,092

Source: World Almanac.

21. Charity Donations A random sample of 30 large companies in the United States shows the amount,

in millions of dollars, that each company donated to charity for a specific year. Construct a frequency distribution for the data, using 9 classes.

26	25	19	31	14
48	35	43	25	46
17	21	57	58	34
41	12	27	15	53
16	63	82	23	52
56	75	19	26	88

22. Unclaimed Expired Prizes The number of unclaimed expired prizes (in millions of dollars) for lottery tickets bought in a sample of states is shown. Construct a frequency distribution for the data, using 5 classes.

28.5	51.7	19	5
2	1.2	14	14.6
0.8	11.6	3.5	30.1
1.7	1.3	13	14

23. Scores in the Rose Bowl The data show the scores of the winning teams in the Rose Bowl. Construct a frequency distribution for the data using a class width of 7.

24	20	45	21	26	38	49	32	41	38
28	34	37	34	17	38	21	20	41	38
21	38	34	46	17	22	20	22	45	20
45	24	28	23	17	17	27	14	23	18

Source: The World Almanac.

24. Consumption of Natural Gas Construct a frequency distribution for the energy consumption of natural gas (in billions of Btu) by the 50 states and the District of Columbia. Use 9 classes.

474	475	205	639	197	344	3	409	247	66
377	87	747	1166	223	248	958	406	251	3462
2391	514	371	58	224	530	317	267	769	9
188	289	76	678	331	52	214	165	255	319
34	1300	284	834	114	1082	73	62	95	393
146									

Source: Time Almanac.

25. Average Wind Speeds A sample of 40 large cities was selected, and the average of the wind speeds was computed for each city over one year. Construct a frequency distribution, using 7 classes.

12.2	9.1	11.2	9.0
10.5	8.2	8.9	12.2
9.5	10.2	7.1	11.0
6.2	7.9	8.7	8.4
8.9	8.8	7.1	10.1
8.7	10.5	10.2	10.7
7.9	8.3	8.7	8.7
10.4	7.7	12.3	10.7
7.7	7.8	11.8	10.5
9.6	9.6	8.6	10.3

Source: World Almanac and Book of Facts.

26. Percentage of People Who Completed 4 or More Years of College Listed by state are the percentages of the population who have completed 4 or more years of a college education. Construct a frequency distribution with 7 classes.

21.4	26.0	25.3	19.3	29.5	35.0	34.7	26.1	25.8	23.4
27.1	29.2	24.5	29.5	22.1	24.3	28.8	20.0	20.4	26.7
35.2	37.9	24.7	31.0	18.9	24.5	27.0	27.5	21.8	32.5
33.9	24.8	31.7	25.6	25.7	24.1	22.8	28.3	25.8	29.8
23.5	25.0	21.8	25.2	28.7	33.6	33.6	30.3	17.3	25.4

Source: New York Times Almanac.

Extending the Concepts

27. JFK Assassination A researcher conducted a survey asking people if they believed more than one person was involved in the assassination of John F. Kennedy. The results were as follows: 73% said yes, 19% said no, and 9% had no opinion. Is there anything suspicious about the results?

28. The Value of Pi The ratio of the circumference of a circle to its diameter is known as π (pi). The value of π is an irrational number, which means that the decimal part goes on forever and there is no fixed sequence of numbers that repeats. People have found the decimal part of π to over a million places. We can statistically study the number. Shown here is the value of π to 40 decimal places. Construct an ungrouped frequency distribution for the digits. Based on the distribution, do you think each digit appears equally in the number?

3.1415926535897932384626433832795028841971

2–2 Histograms, Frequency Polygons, and Ogives

OBJECTIVE 2

Represent data in frequency distributions graphically, using histograms, frequency polygons, and ogives.

After you have organized the data into a frequency distribution, you can present them in graphical form. The purpose of graphs in statistics is to convey the data to the viewers in pictorial form. It is easier for most people to comprehend the meaning of data presented graphically than data presented numerically in tables or frequency distributions. This is especially true if the users have little or no statistical knowledge.

Statistical graphs can be used to describe the data set or to analyze it. Graphs are also useful in getting the audience's attention in a publication or a speaking presentation. They can be used to discuss an issue, reinforce a critical point, or summarize a data set. They can also be used to discover a trend or pattern in a situation over a period of time.

The three most commonly used graphs in research are

1. The histogram.
2. The frequency polygon.
3. The cumulative frequency graph, or ogive (pronounced o-jive).

The steps for constructing the histogram, frequency polygon, and the ogive are summarized in the procedure table.

Procedure Table

Constructing a Histogram, Frequency Polygon, and Ogive

Step 1	Draw and label the x and y axes.
Step 2	On the x axis, label the class boundaries of the frequency distribution for the histogram and ogive. Label the midpoints for the frequency polygon.
Step 3	Plot the frequencies for each class, and draw the vertical bars for the histogram and the lines for the frequency polygon and ogive.

(*Note:* Remember that the lines for the frequency polygon begin and end on the x axis while the lines for the ogive begin on the x axis.)

Historical Note

Karl Pearson introduced the histogram in 1891. He used it to show time concepts of various reigns of Prime Ministers.

The Histogram

The **histogram** is a graph that displays the data by using contiguous vertical bars (unless the frequency of a class is 0) of various heights to represent the frequencies of the classes.

EXAMPLE 2–4 Record High Temperatures

Construct a histogram to represent the data shown for the record high temperatures for each of the 50 states (see Example 2–2).

Class boundaries	Frequency
99.5–104.5	2
104.5–109.5	8
109.5–114.5	18
114.5–119.5	13
119.5–124.5	7
124.5–129.5	1
129.5–134.5	1

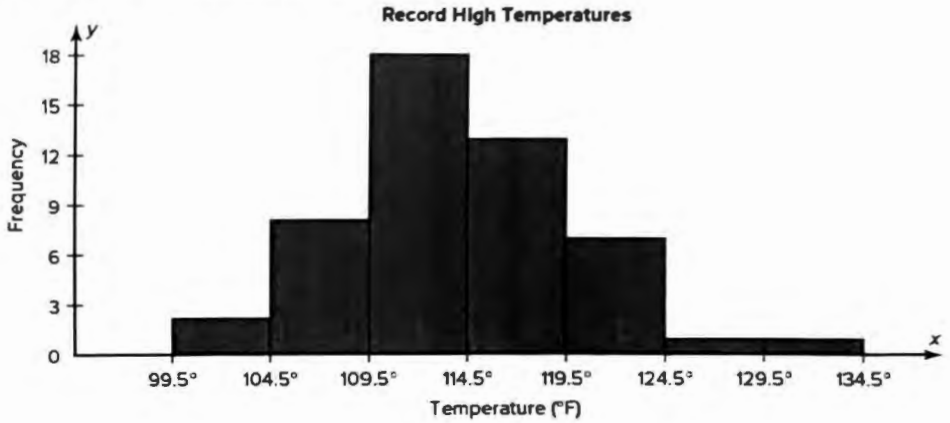
SOLUTION

Step 1 Draw and label the x and y axes. The x axis is always the horizontal axis, and the y axis is always the vertical axis.

Step 2 Represent the frequency on the y axis and the class boundaries on the x axis.

Step 3 Using the frequencies as the heights, draw vertical bars for each class. See Figure 2–1.

FIGURE 2–1 Histogram for Example 2–4



As the histogram shows, the class with the greatest number of data values (18) is 109.5–114.5, followed by 13 for 114.5–119.5. The graph also has one peak with the data clustering around it.

Historical Note

Graphs originated when ancient astronomers drew the position of the stars in the heavens. Roman surveyors also used coordinates to locate landmarks on their maps.

The development of statistical graphs can be traced to William Playfair (1759–1823), an engineer and drafter who used graphs to present economic data pictorially.

The Frequency Polygon

Another way to represent the same data set is by using a frequency polygon.

The frequency polygon is a graph that displays the data by using lines that connect points plotted for the frequencies at the midpoints of the classes. The frequencies are represented by the heights of the points.

Example 2–5 shows the procedure for constructing a frequency polygon. Be sure to begin and end on the x axis.

EXAMPLE 2–5 Record High Temperatures

Using the frequency distribution given in Example 2–4, construct a frequency polygon.

SOLUTION

Step 1 Find the midpoints of each class. Recall that midpoints are found by adding the upper and lower boundaries and dividing by 2:

$$\frac{99.5 + 104.5}{2} = 102 \qquad \frac{104.5 + 109.5}{2} = 107$$

and so on. The midpoints are

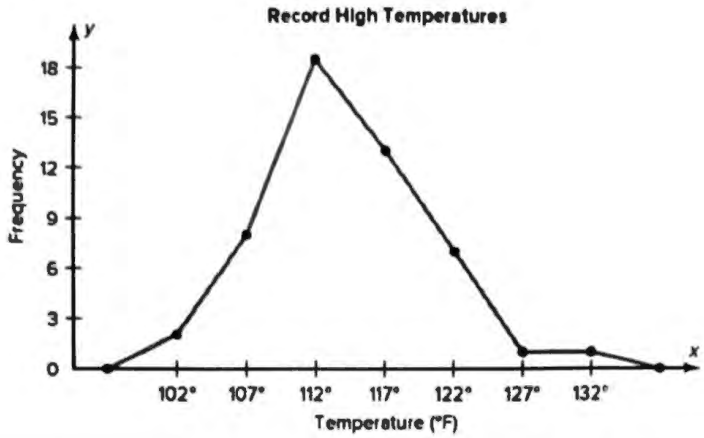
Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

Step 2 Draw the x and y axes. Label the x axis with the midpoint of each class, and then use a suitable scale on the y axis for the frequencies.

Step 3 Using the midpoints for the x values and the frequencies as the y values, plot the points.

Step 4 Connect adjacent points with line segments. Draw a line back to the x axis at the beginning and end of the graph, at the same distance that the previous and next midpoints would be located, as shown in Figure 2–2.

FIGURE 2–2
Frequency Polygon for Example 2–5



The frequency polygon and the histogram are two different ways to represent the same data set. The choice of which one to use is left to the discretion of the researcher.

The Ogive

The third type of graph that can be used represents the cumulative frequencies for the classes. This type of graph is called the *cumulative frequency graph*, or *ogive*. The **cumulative frequency** is the sum of the frequencies accumulated up to the upper boundary of a class in the distribution.

The ogive is a graph that represents the cumulative frequencies for the classes in a frequency distribution.

Example 2–6 shows the procedure for constructing an ogive. Be sure to start on the x axis.

EXAMPLE 2–6 Record High Temperatures

Construct an ogive for the frequency distribution described in Example 2–4.

SOLUTION

Step 1 Find the cumulative frequency for each class.

	Cumulative frequency
Less than 99.5	0
Less than 104.5	2
Less than 109.5	10
Less than 114.5	28
Less than 119.5	41
Less than 124.5	48
Less than 129.5	49
Less than 134.5	50

Step 2 Draw the x and y axes. Label the x axis with the class boundaries. Use an appropriate scale for the y axis to represent the cumulative frequencies. (Depending on the numbers in the cumulative frequency columns, scales such as 0, 1, 2, 3, . . . , or 5, 10, 15, 20, . . . , or 1000, 2000, 3000, . . . can be used. Do *not* label the y axis with the numbers in the cumulative frequency column.) In this example, a scale of 0, 5, 10, 15, . . . will be used.

Step 3 Plot the cumulative frequency at each upper class boundary, as shown in Figure 2–3. Upper boundaries are used since the cumulative frequencies represent the number of data values accumulated up to the upper boundary of each class.

Step 4 Starting with the first upper class boundary, 104.5, connect adjacent points with line segments, as shown in Figure 2–4. Then extend the graph to the first lower class boundary, 99.5, on the x axis.

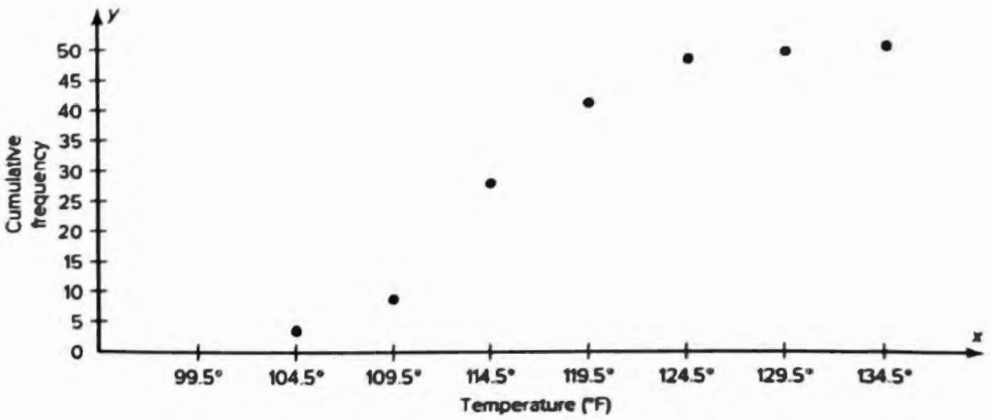


FIGURE 2–3
Plotting the Cumulative Frequency for Example 2–6

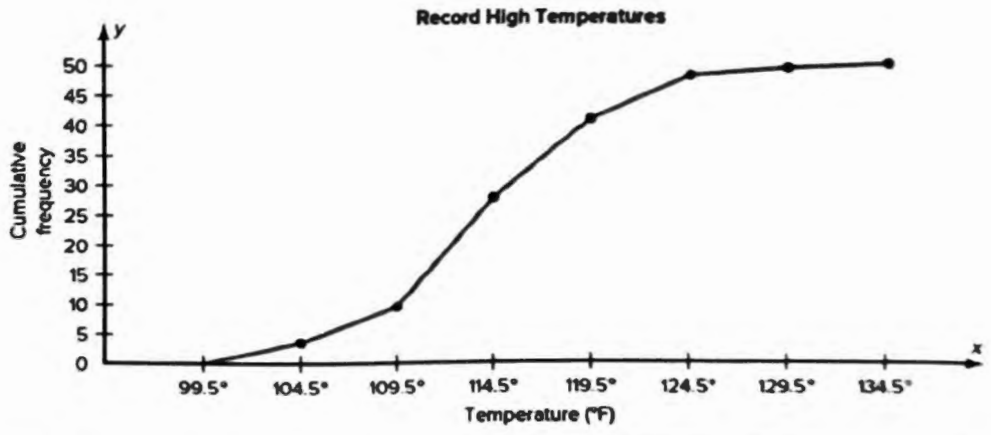
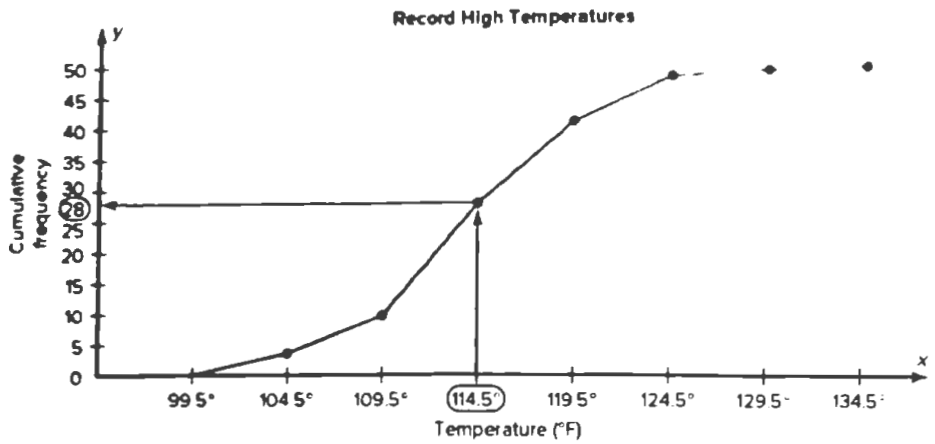


FIGURE 2–4
Ogive for Example 2–6

FIGURE 2–5
Finding a Specific Cumulative Frequency



Unusual Stats
Twenty-two percent of Americans sleep 6 hours a day or less

Cumulative frequency graphs are used to visually represent how many values are below a certain upper class boundary. For example, to find out how many record high temperatures are less than 114.5°F, locate 114.5°F on the x axis, draw a vertical line up until it intersects the graph, and then draw a horizontal line at that point to the y axis. The y axis value is 28, as shown in Figure 2–5.

Relative Frequency Graphs

The histogram, the frequency polygon, and the ogive shown previously were constructed by using frequencies in terms of the raw data. These distributions can be converted to distributions using *proportions* instead of raw data as frequencies. These types of graphs are called **relative frequency graphs.**

Graphs of relative frequencies instead of frequencies are used when the proportion of data values that fall into a given class is more important than the actual number of data values that fall into that class. For example, if you wanted to compare the age distribution of adults in Philadelphia, Pennsylvania, with the age distribution of adults of Erie, Pennsylvania, you would use relative frequency distributions. The reason is that since the population of Philadelphia is 1,526,006 and the population of Erie is 101,786, the bars using the actual data values for Philadelphia would be much taller than those for the same classes for Erie.

To convert a frequency into a proportion or relative frequency, divide the frequency for each class by the total of the frequencies. The sum of the relative frequencies will always be 1. These graphs are similar to the ones that use raw data as frequencies, but the values on the y axis are in terms of proportions. Example 2–7 shows the three types of relative frequency graphs.

EXAMPLE 2–7 Ages of State Governors

Construct a histogram, frequency polygon, and ogive using relative frequencies for the distribution shown. This is a grouped frequency distribution using the ages (at the time of this writing) of the governors of the 50 states of the United States.

42.5–47.5	4
47.5–52.5	4
52.5–57.5	11
57.5–62.5	14
62.5–67.5	9
67.5–72.5	5
72.5–77.5	3
	Total 50

SOLUTION

Step 1 Convert each frequency to a proportion or relative frequency by dividing the frequency for each class by the total number of observations.

For the class 42.5–47.5 the relative frequency $= \frac{4}{50} = 0.08$; for the class 47.5–52.5, the relative frequency is $\frac{4}{50} = 0.08$; for the class 52.5–57.5, the relative frequency is $\frac{11}{50} = 0.22$, and so on.

Place these values in the column labeled Relative Frequency. Also, find the midpoints, as shown in Example 2-5, for each class and place them in the midpoint column

Class boundaries	Midpoints	Relative frequency
42.5–47.5	45	0.08
47.5–52.5	50	0.08
52.5–57.5	55	0.22
57.5–62.5	60	0.28
62.5–67.5	65	0.18
67.5–72.5	70	0.10
72.5–77.5	75	0.06

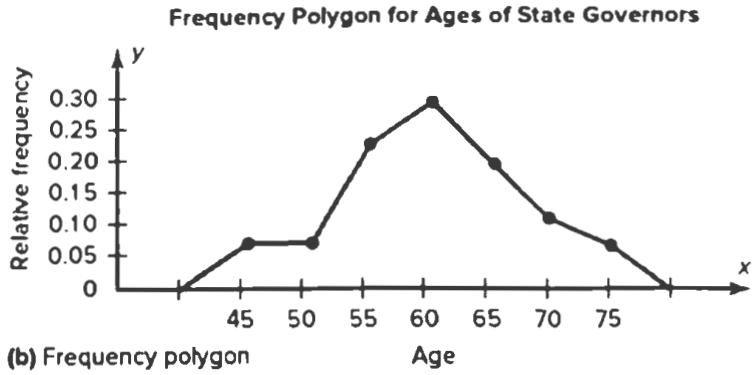
Step 2 Find the cumulative relative frequencies. To do this, add the frequency in each class to the total frequency of the preceding class. In this case, 0.00 + 0.08 = 0.08, 0.08 + 0.08 = 0.16, 0.16 + 0.22 = 0.38, 0.28 + 0.38 = 0.66, etc. Place these values in a column labeled Cumulative relative frequency.

An alternative method would be to change the cumulative frequencies for the classes to relative frequencies. (Divide each by the total).

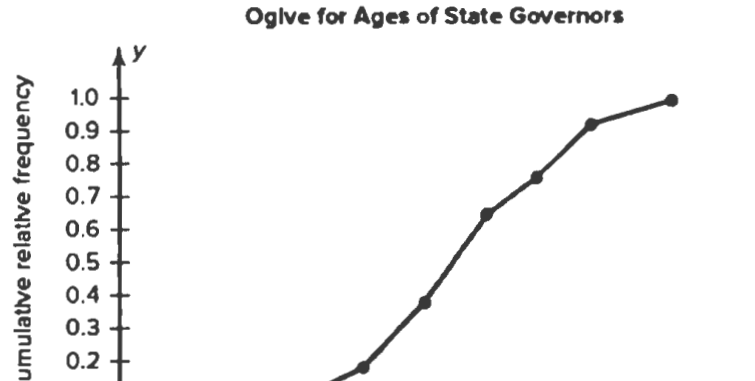
	Cumulative frequency	Cumulative relative frequency
Less than 42.5	0	0.00
Less than 47.5	4	0.08
Less than 52.5	8	0.16
Less than 57.5	19	0.38
Less than 62.5	33	0.66
Less than 67.5	42	0.84
Less than 72.5	47	0.94
Less than 77.5	50	1.00

Step 3 Draw each graph as shown in Figure 2-6. For the histogram and ogive, use the class boundaries along the x axis. For the frequency, use the midpoints on the x axis. For the scale on the y axis, use proportions.

FIGURE 2–6
Graphs for Example 2–7



(b) Frequency polygon



(c) Ogive

Distribution Shapes

When one is describing data, it is important to be able to recognize the shapes of the distribution values. In later chapters, you will see that the shape of a distribution also determines the appropriate statistical methods used to analyze the data.

A distribution can have many shapes, and one method of analyzing a distribution is to draw a histogram or frequency polygon for the distribution. Several of the most common shapes are shown in Figure 2–7: *the bell-shaped or mound-shaped, the uniform-shaped, the J-shaped, the reverse J-shaped, the positively or right-skewed shape, the negatively or left-skewed shape, the bimodal-shaped, and the U-shaped.*

Distributions are most often not perfectly shaped, so it is not necessary to have an exact shape but rather to identify an overall pattern.

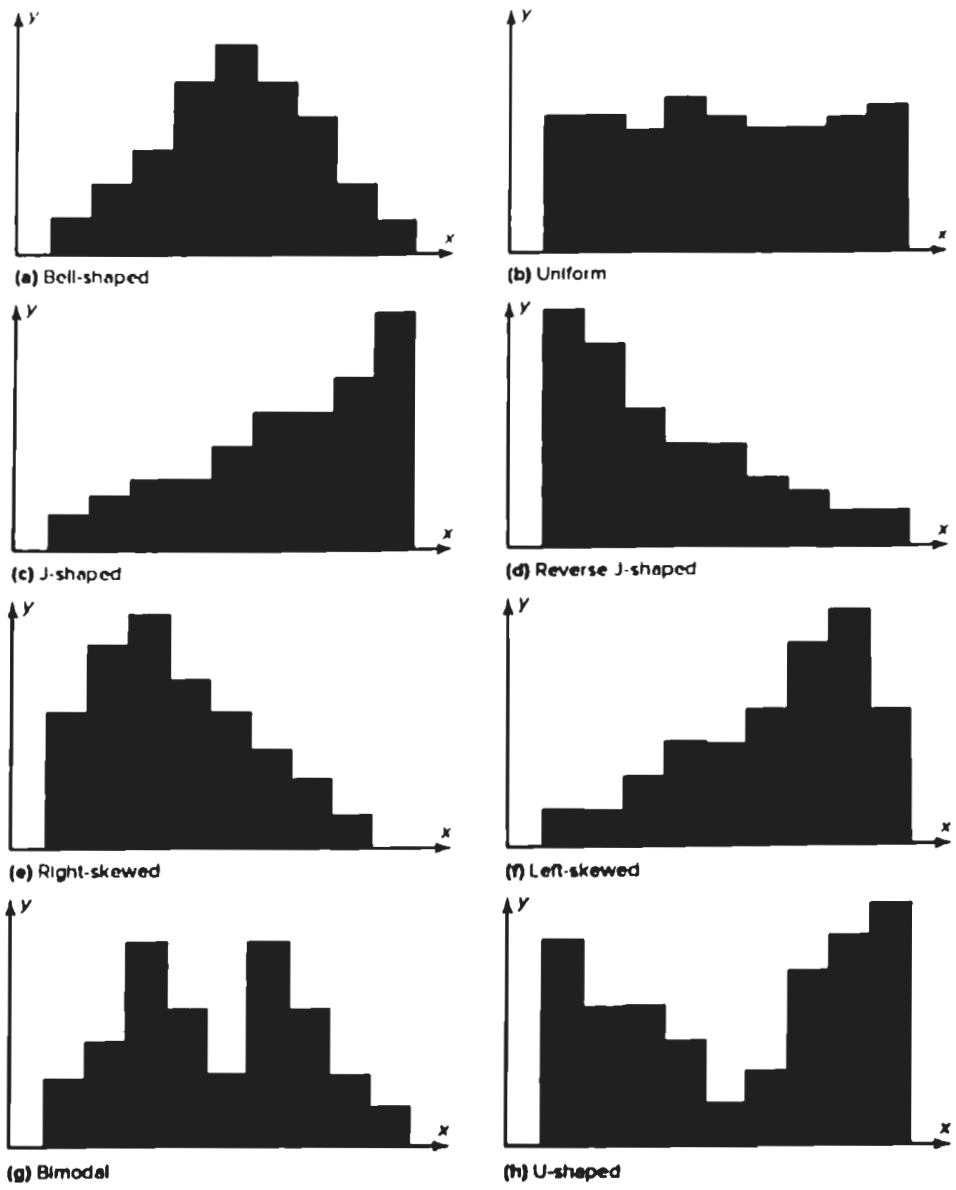
A *bell-shaped distribution* shown in Figure 2–7(a) has a single peak and tapers off at either end. It is approximately symmetric; i.e., it is roughly the same on both sides of a line running through the center.

A *uniform distribution* is basically flat or rectangular. See Figure 2–7(b).

A *J-shaped distribution* is shown in Figure 2–7(c), and it has a few data values on the left side and increases as one moves to the right. A *reverse J-shaped distribution* is the opposite of the J-shaped distribution. See Figure 2–7(d).

When the peak of a distribution is to the left and the data values taper off to the right, a distribution is said to be *positively or right-skewed.* See Figure 2–7(e). When

FIGURE 2–7
Distribution Shapes



the data values are clustered to the right and taper off to the left, a distribution is said to be *negatively or left-skewed.* See Figure 2–7(f). Skewness will be explained in detail in Chapter 3. Distributions with one peak, such as those shown in Figure 2–7(a), (e), and (f), are said to be *unimodal.* (The highest peak of a distribution indicates where the mode of the data values is. The mode is the data value that occurs more often than any other data value. Modes are explained in Chapter 3.) When a distribution has two peaks of the same height, it is said to be *bimodal.* See Figure 2–7(g). Finally, the graph shown in Figure 2–7(h) is a *U-shaped* distribution.

Distributions can have other shapes in addition to the ones shown here; however, these are some of the more common ones that you will encounter in analyzing data.

When you are analyzing histograms and frequency polygons, look at the shape of the curve. For example, does it have one peak or two peaks? Is it relatively flat, or is it U-shaped? Are the data values spread out on the graph, or are they clustered around the center? Are there data values in the extreme ends? These may be *outliers.* (See Section 3–3 for an explanation of outliers.) Are there any gaps in the histogram, or does

the frequency polygon touch the x axis somewhere other than at the ends? Finally, are the data clustered at one end or the other, indicating a *skewed distribution*?

For example, the histogram for the record high temperatures in Figure 2–1 shows a single peaked distribution, with the class 109.5–114.5 containing the largest number of temperatures. The distribution has no gaps, and there are fewer temperatures in the highest class than in the lowest class.

Applying the Concepts 2–2

Selling Real Estate

Assume you are a realtor in Bradenton, Florida. You have recently obtained a listing of the selling prices of the homes that have sold in that area in the last 6 months. You wish to organize those data so you will be able to provide potential buyers with useful information. Use the following data to create a histogram, frequency polygon, and cumulative frequency polygon.

142,000	127,000	99,600	162,000	89,000	93,000	99,500
73,800	135,000	119,500	67,900	156,300	104,500	108,650
123,000	91,000	205,000	110,000	156,300	104,000	133,900
179,000	112,000	147,000	321,550	87,900	88,400	180,000
159,400	205,300	144,400	163,000	96,000	81,000	131,000
114,000	119,600	93,000	123,000	187,000	96,000	80,000
231,000	189,500	177,600	83,400	77,000	132,300	166,000

1. What questions could be answered more easily by looking at the histogram rather than the listing of home prices?
2. What different questions could be answered more easily by looking at the frequency polygon rather than the listing of home prices?
3. What different questions could be answered more easily by looking at the cumulative frequency polygon rather than the listing of home prices?
4. Are there any extremely large or extremely small data values compared to the other data values?
5. Which graph displays these extremes the best?
6. Is the distribution skewed?

See page 108 for the answers.

Exercises 2–2

1. **Do Students Need Summer Development?** For 108 randomly selected college applicants, the following frequency distribution for entrance exam scores was obtained. Construct a histogram, frequency polygon, and ogive for the data. (The data for this exercise will be used for Exercise 13 in this section.)

Class limits	Frequency
90–98	6
99–107	22
108–116	43
117–125	28
126–134	9
	Total 108

Applicants who score above 107 need not enroll in a summer developmental program. In this group, how many students do not have to enroll in the developmental program?

2. **Bear Kills** The number of bears killed in 2014 for 56 counties in Pennsylvania is shown in the frequency distribution. Construct a histogram, frequency polygon, and ogive for the data. Comment on the skewness of the distribution. How many counties had 75 or fewer bears killed? (The data for this exercise will be used for Exercise 14 of this section.)

Class limits	Frequency
1–25	16
26–50	14
51–75	9
76–100	8
101–125	5
126–150	0
151–175	1
176–200	1
201–225	0
226–250	0
251–275	2
	Total 56

Source: Pennsylvania State Game Commission.

3. **Pupils Per Teacher** The average number of pupils per teacher in each state is shown. Construct a grouped frequency distribution with 6 classes. Draw a histogram, frequency polygon, and ogive. Analyze the distribution.

16	16	15	12	14
13	16	14	15	14
18	18	18	12	15
15	16	16	15	15
25	19	15	12	22
18	14	13	17	9
13	14	13	16	12
14	16	10	22	20
12	14	18	15	14
16	12	12	13	15

Source: U.S. Department of Education.

4. **Number of College Faculty** The number of faculty listed for a sample of private colleges that offer only bachelor's degrees is listed below. Use these data to construct a frequency distribution with 7 classes, a histogram, a frequency polygon, and an ogive. Discuss the shape of this distribution. What proportion of schools have 180 or more faculty?

165	221	218	206	138	135	224	204
70	210	207	154	155	82	120	116
176	162	225	214	93	389	77	135
221	161	128	310				

Source: World Almanac and Book of Facts.

5. **Railroad Crossing Accidents** The data show the number of railroad crossing accidents for the 50 states of the United States for a specific year. Construct a histogram, frequency polygon, and ogive for the data. Comment on the skewness of the distribution. (The data in this exercise will be used for Exercise 15 in this section.)

Class limits	Frequency
1–43	24
44–86	17
87–129	3
130–172	4
173–215	1
216–258	0
259–301	0
302–344	1
	Total 50

Source: Federal Railroad Administration.

6. **NFL Salaries** The salaries (in millions of dollars) for 31 NFL teams for a specific season are given in this frequency distribution.

Construct a histogram, a frequency polygon, and an ogive for the data; and comment on the shape of the distribution. (The data for this exercise will be used for Exercise 16 of this section.)

Class limits	Frequency
39.9–42.8	2
42.9–45.8	2
45.9–48.8	5
48.9–51.8	5
51.9–54.8	12
54.9–57.8	5
	Total 31

Source: NFL.com

7. **Suspension Bridges Spans** The following frequency distribution shows the length (in feet) of the main spans of the longest suspension bridges in the United States. Construct a histogram, frequency polygon, and ogive for the distribution. Describe the shape of the distribution.

Class limits	Frequency
1260–1734	12
1735–2209	6
2210–2684	3
2685–3159	1
3160–3634	1
3635–4109	1
4110–4584	2

Source: U.S. Department of Transportation.

8. **Costs of Utilities** The frequency distribution represents the cost (in cents) for the utilities of states that supply much of their own power. Construct a histogram, frequency polygon, and ogive for the data. Is the distribution skewed?

Class limits	Frequency
6–8	12
9–11	16
12–14	3
15–17	1
18–20	0
21–23	0
24–26	1
	Total 33

9. **Air Pollution** One of the air pollutants that is measured in selected cities is sulfur dioxide. This pollutant occurs when fossil fuels are burned. This pollutant is measured in micrograms per cubic meter ($\mu g/m^3$). The results obtained from a sample of 24 cities are shown in the frequency distributions. One sample was taken recently, and the other sample of the same cities was taken

5 years ago. Construct a histogram and compare the two distributions.

Class limits	Frequency (now)	Frequency (5 years ago)
10–14	6	5
15–19	4	4
20–24	3	2
25–29	2	3
30–34	5	6
35–39	1	2
40–44	2	1
45–49	1	1
	Total 24	Total 24

10. **Making the Grade** The frequency distributions shown indicate the percentages of public school students in fourth-grade reading and mathematics who performed at or above the required proficiency levels for the 50 states in the United States. Draw histograms for each, and decide if there is any difference in the performance of the students in the subjects.

Class	Reading frequency	Math frequency
17.5–22.5	7	5
22.5–27.5	6	9
27.5–32.5	14	11
32.5–37.5	19	16
37.5–42.5	3	8
42.5–47.5	1	1
	Total 50	Total 50

Source: National Center for Educational Statistics

11. **Blood Glucose Levels** The frequency distribution shows the blood glucose levels (in milligrams per deciliter) for 50 patients at a medical facility. Construct a histogram, frequency polygon, and ogive for the data. Comment on the shape of the distribution. What range of glucose levels did most patients fall into?

Class limits	Frequency
60–64	2
65–69	1
70–74	5
75–79	12
80–84	18
85–89	6
90–94	5
95–99	1
	Total 50

12. **Waiting Times** The frequency distribution shows the waiting times (in minutes) for 50 patients at a walk-in medical facility. Construct a histogram, frequency polygon, and ogive for the data. Is the distribution skewed? How many patients waited longer than 30 minutes?

Class limits	Frequency
11–15	7
16–20	9
21–25	15
26–30	9
31–35	5
36–40	3
41–45	2
	Total 50

13. Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 1 of this section.

14. Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 2 of this section.

15. Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 5 of this section.

16. Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 6 of this section.

17. **Home Runs** The data show the most number of home runs hit by a batter in the American League over the last 30 seasons. Construct a frequency distribution using 5 classes. Draw a histogram, a frequency polygon, and an ogive for the date, using relative frequencies. Describe the shape of the histogram.

40	43	40
53	47	46
44	57	43
43	52	44
54	47	51
39	48	36
37	56	42
54	56	49
54	52	40
48	50	40

Source: World Almanac and Book of Facts.

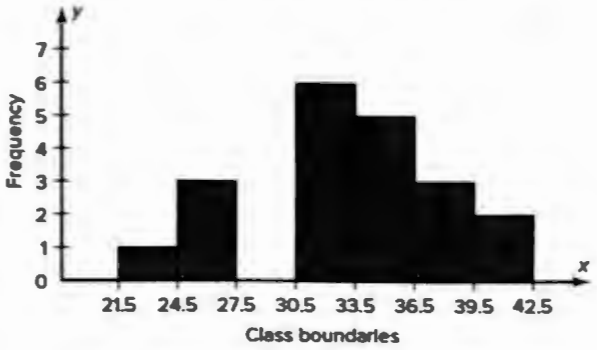
18. **Protein Grams in Fast Food** The amount of protein (in grams) for a variety of fast-food sandwiches is reported here. Construct a frequency distribution, using 6 classes. Draw a histogram, a frequency polygon, and an ogive for the data, using relative frequencies. Describe the shape of the histogram.

23	30	20	27	44	26	35	20	29	29
25	15	18	27	19	22	12	26	34	15
27	35	26	43	35	14	24	12	23	31
40	35	38	57	22	42	24	21	27	33

Source: The Doctor's Pocket Calorie, Fat, and Carbohydrate Counter.

Extending the Concepts

19. Using the histogram shown here, do the following.



a. Construct a frequency distribution; include class limits, class frequencies, midpoints, and cumulative frequencies.

b. Construct a frequency polygon.

c. Construct an ogive.

20. Using the results from Exercise 19, answer these questions.

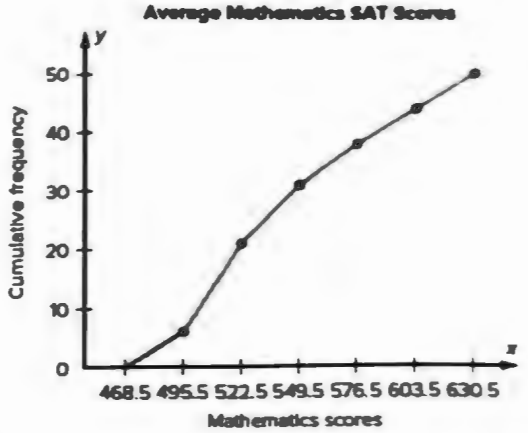
a. How many values are in the class 27.5–30.5?

b. How many values fall between 24.5 and 36.5?

c. How many values are below 33.5?

d. How many values are above 30.5?

21. Math SAT Scores Shown is an ogive depicting the cumulative frequency of the average mathematics SAT scores by state. Use it to construct a histogram and a frequency polygon.



Technology Step by Step

TI-84 Plus

Step by Step

Constructing a Histogram

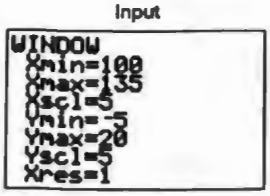
To display the graphs on the screen, enter the appropriate values in the calculator, using the WINDOW menu. The default values are $X_{min} = -10$, $X_{max} = 10$, $Y_{min} = -10$, and $Y_{max} = 10$.

The X_{scl} changes the distance between the tick marks on the x axis and can be used to change the class width for the histogram.

To change the values in the WINDOW:

1. Press **WINDOW**.
2. Move the cursor to the value that needs to be changed. Then type in the desired value and press **ENTER**.
3. Continue until all values are appropriate.
4. Press **[2nd] [QUIT]** to leave the WINDOW menu.

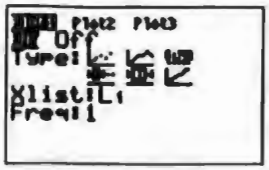
Input



To plot the histogram from raw data:

1. Enter the data in L_1.
2. Make sure WINDOW values are appropriate for the histogram.
3. Press **[2nd] [STAT PLOT] ENTER**.
4. Press **ENTER** to turn the plot 1 on, if necessary.
5. Move cursor to the Histogram symbol and press **ENTER**, if necessary. The histogram is the third option.
6. Make sure Xlist is L_1.
7. Make sure Freq is 1.
8. Press **GRAPH** to display the histogram.
9. To obtain the frequency (number of data values in each class), press the **TRACE** key, followed by ◀ or ▶ keys.

Input



Output

Example TI2–1

Plot a histogram for the following data from Example 2–2.

112	100	127	120	134	118	105	110	109	112
110	118	117	116	118	122	114	114	105	109
107	112	114	115	118	117	118	122	106	110
116	108	110	121	113	120	119	111	104	111
120	113	120	117	105	110	118	112	114	114

Press **TRACE** and use the arrow keys to determine the number of values in each group.

To graph a histogram from grouped data:

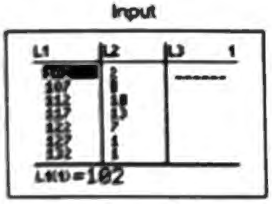
1. Enter the midpoints into L_1.
2. Enter the frequencies into L_2.
3. Make sure WINDOW values are appropriate for the histogram.
4. Press **[2nd] [STAT PLOT] ENTER.**
5. Press **ENTER** to turn the plot on, if necessary.
6. Move cursor to the histogram symbol, and press **ENTER,** if necessary.
7. Make sure Xlist is L_1.
8. Make sure Freq is L_2.
9. Press **GRAPH** to display the histogram.

Example TI2–2

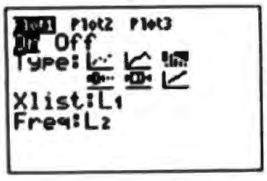
Plot a histogram for the data from Examples 2–4 and 2–5.

Class boundaries	Midpoints	Frequency
99.5–104.5	102	2
104.5–109.5	107	8
109.5–114.5	112	18
114.5–119.5	117	13
119.5–124.5	122	7
124.5–129.5	127	1
129.5–134.5	132	1

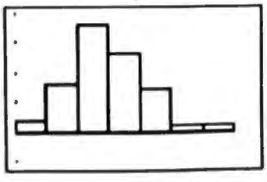
Input



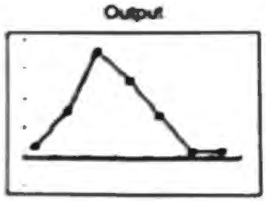
Input



Output



Output



To graph a frequency polygon from grouped data, follow the same steps as for the histogram except change the graph type from histogram (third graph) to a line graph (second graph).

Output

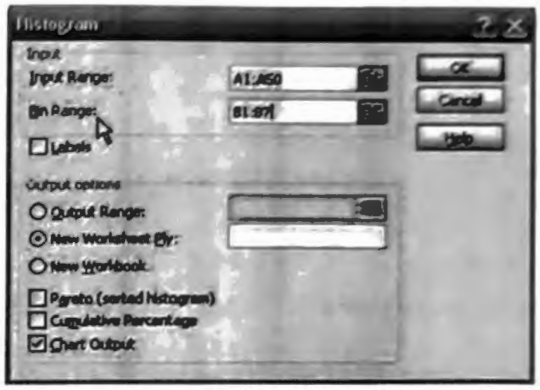
To graph an ogive from grouped data, modify the procedure for the histogram as follows:

1. Enter the upper class boundaries into L_1.
2. Enter the cumulative frequencies into L_2.
3. Change the graph type from histogram (third graph) to line (second graph).
4. Change the Y_{max} from the WINDOW menu to the sample size.

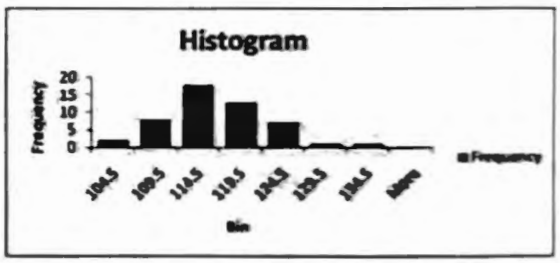
EXCEL
Step by Step

Constructing a Histogram

1. Press [Ctrl]-N for a new workbook.
2. Enter the data from Example 2–2 in column A, one number per cell.
3. Enter the upper boundaries into column B.
4. From the toolbar, select the Data tab, then select Data Analysis.
5. In Data Analysis, select Histogram and click [OK].
6. In the Histogram dialog box, type **A1:A50** in the Input Range box and type **B1:B7** in the Bin Range box.



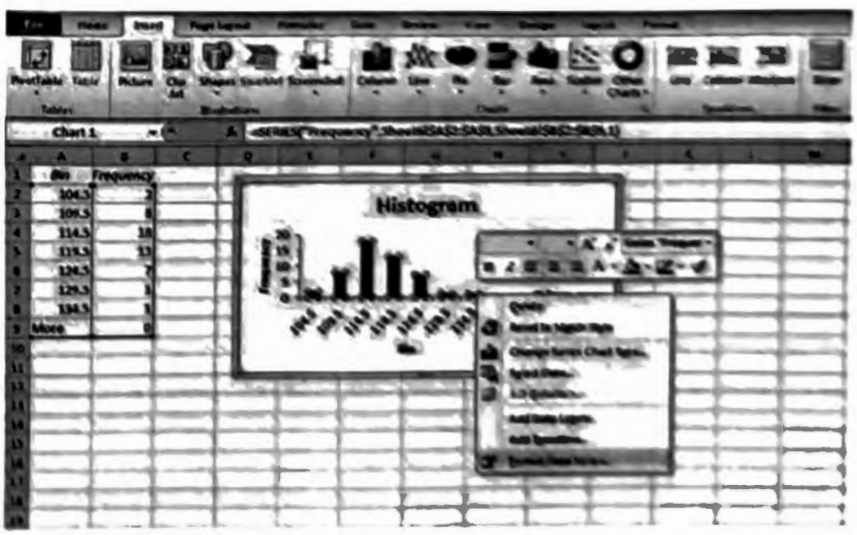
7. Select New Worksheet Ply and Chart Output. Click [OK].



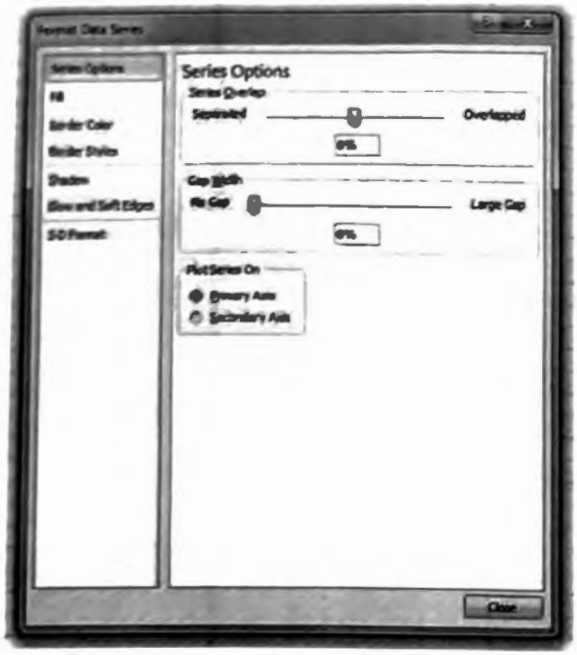
Editing the Histogram

To move the vertical bars of the histogram closer together:

1. Right-click one of the bars of the histogram, and select Format Data Series.

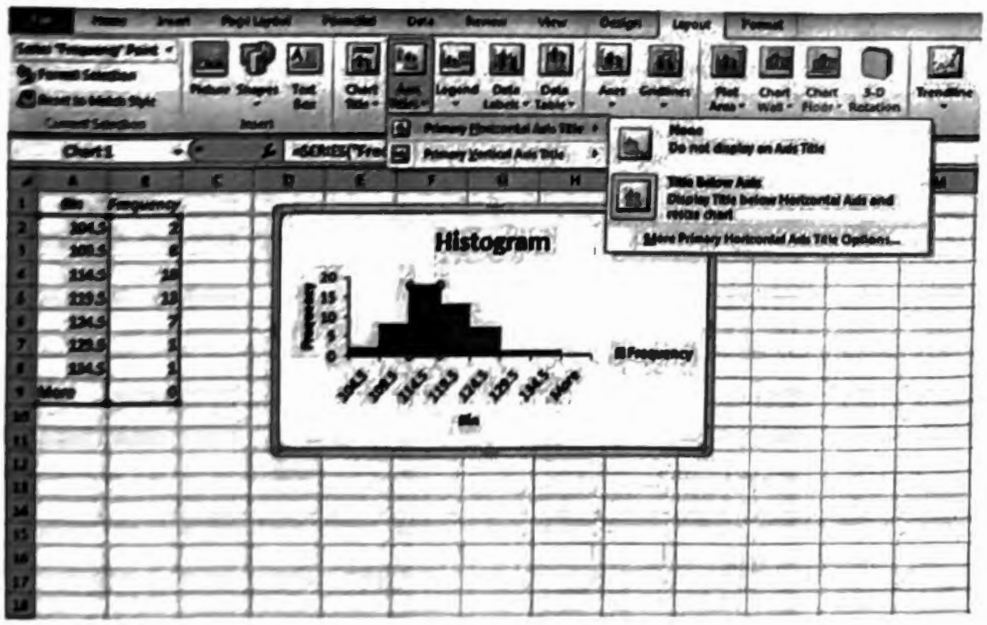


2. Move the Gap Width slider all the way to the left to change the gap width of the bars in the histogram to 0.



To change the label for the horizontal axis:

1. Left-click the mouse over any part of the histogram.
2. Select the Chart Tools tab from the toolbar.
3. Select the Layout tab, Axis Titles and Primary Horizontal Axis Title.



Once the Axis Titles text box is selected, you can type in the name of the variable represented on the horizontal axis.

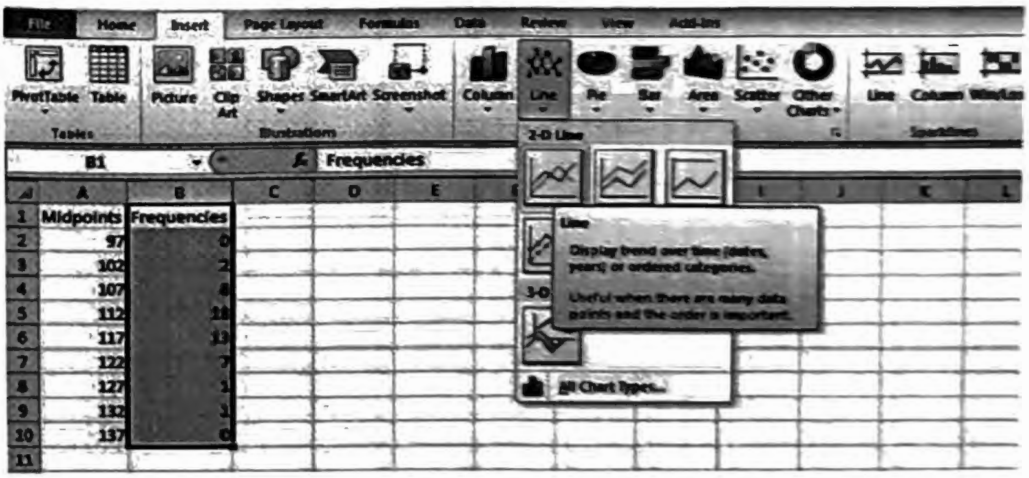
Constructing a Frequency Polygon

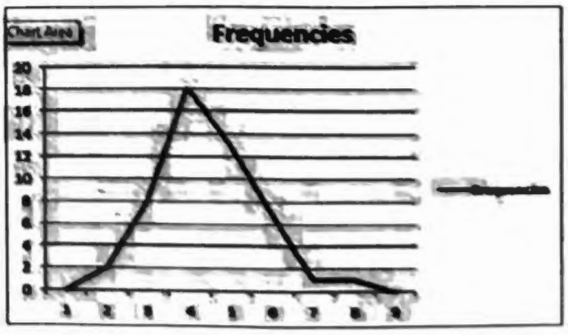
1. Press [CTRL]-N for a new notebook.
2. Enter the midpoints of the data from Example 2–2 into column A and the frequencies into column B, including labels.

Note: Classes with frequency 0 have been added at the beginning and the end to "anchor" the frequency polygon to the horizontal axis.

	A	B
1	Midpoints	Frequencies
2	97	0
3	102	2
4	107	8
5	112	18
6	117	13
7	122	7
8	127	1
9	132	1
10	137	0
11		

3. Press and hold the left mouse button, and drag over the Frequencies (including the label) from column B.
4. Select the Insert tab from the toolbar and the Line Chart option.
5. Select the 2-D line chart type.





We will need to edit the graph so that the midpoints are on the horizontal axis.

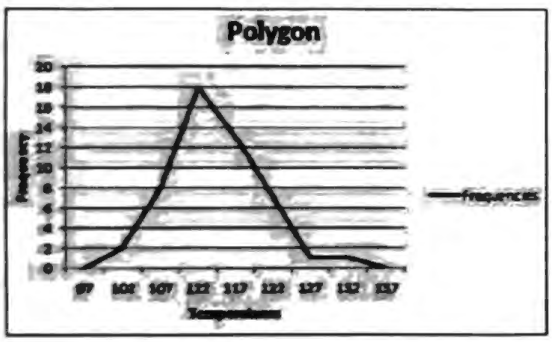
1. Right click the mouse on any region of the chart.
2. Choose Select Data.
3. Select Edit below the Horizontal (Category) Axis Labels panel on the right.
4. Press and hold the left mouse button, and drag over the midpoints (not including the label) for the Axis label range, then click [OK].
5. Click [OK] on the Select Data Source box.

Inserting Labels on the Axes

1. Click the mouse on any region of the graph.
2. Select Chart Tools and then Layout on the toolbar.
3. Select Axis Titles to open the horizontal and vertical axis text boxes. Then manually type in labels for the axes.

Changing the Title

1. Select Chart Tools, Layout from the toolbar.
2. Select Chart Title.
3. Choose one of the options from the Chart Title menu and edit.

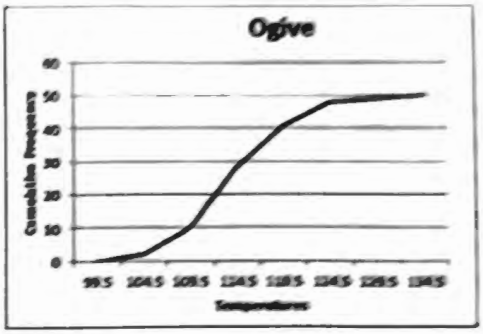


Constructing an Ogive

To create an ogive, use the upper class boundaries (horizontal axis) and cumulative frequencies (vertical axis) from the frequency distribution.

1. Type the upper class boundaries (including a class with frequency 0 before the lowest class to anchor the graph to the horizontal axis) and corresponding cumulative frequencies into adjacent columns of an Excel worksheet.
2. Press and hold the left mouse button, and drag over the Cumulative Frequencies from column B.
3. Select Line Chart, then the 2-D Line option.

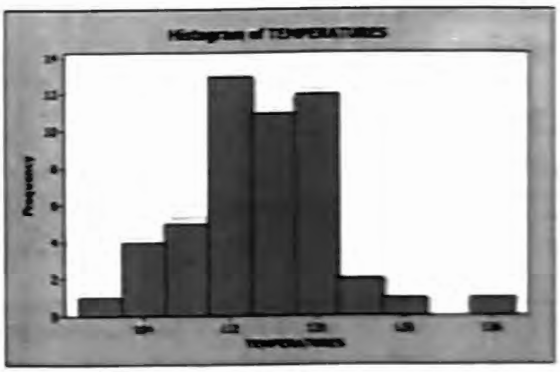
As with the frequency polygon, you can insert labels on the axes and a chart title for the ogive.



MINITAB Step by Step

Construct a Histogram

1. Enter the data from Example 2–2, the high temperatures for the 50 states, into C1.
2. Select **Graph>Histogram**.
3. Select [Simple], then click [OK].
4. Click C1 TEMPERATURES in the Graph variables dialog box.
5. Click [OK]. A new graph window containing the histogram will open.
6. Click the **File** menu to print or save the graph.

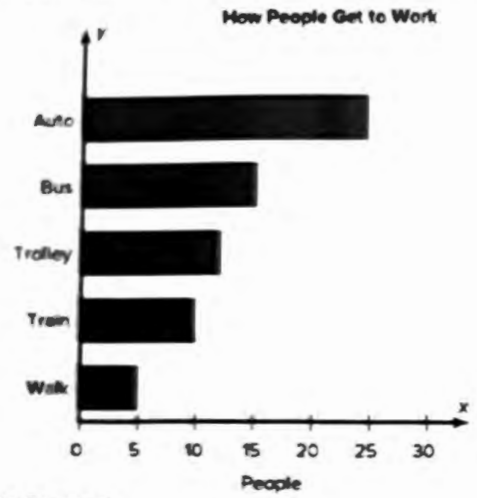


7. Click **File>Exit**.
8. Save the project as **Ch2-3.mpj**. ■

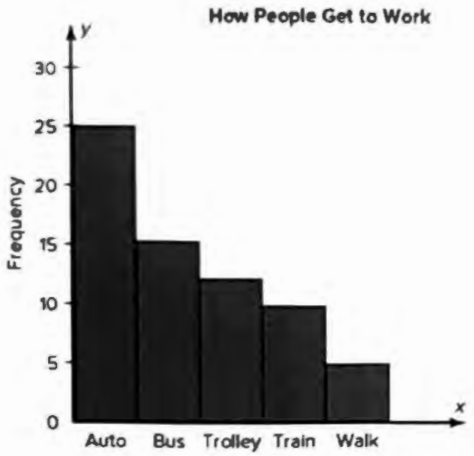
2–3 Other Types of Graphs

In addition to the histogram, the frequency polygon, and the ogive, several other types of graphs are often used in statistics. They are the bar graph, Pareto chart, time series graph, pie graph, and the dotplot. Figure 2–8 shows an example of each type of graph.

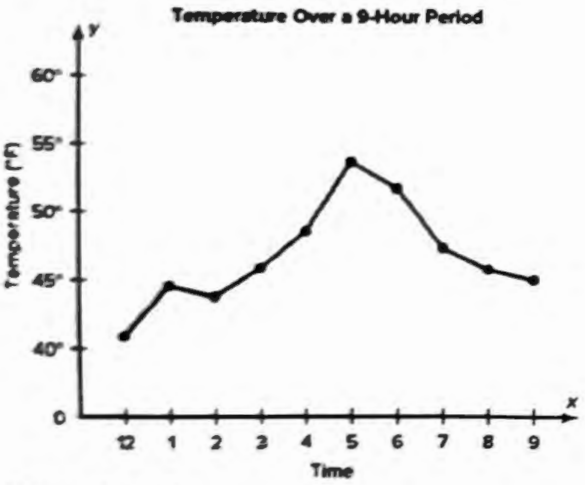
FIGURE 2–8 Other Types of Graphs Used in Statistics



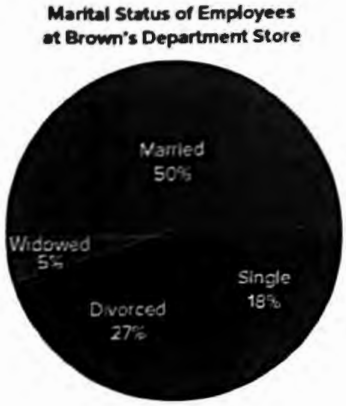
(a) Bar graph



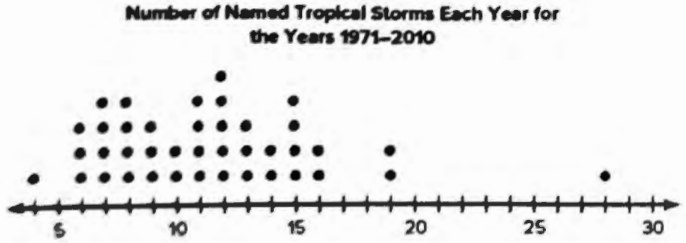
(b) Pareto chart



(c) Time series graph



(d) Pie graph



(e) Dotplot

OBJECTIVE 2

Represent data using bar graphs, Pareto charts, time series graphs, pie graphs, and dotplots.

Bar Graphs

When the data are qualitative or categorical, bar graphs can be used to represent the data. A bar graph can be drawn using either horizontal or vertical bars.

A bar graph represents the data by using vertical or horizontal bars whose heights or lengths represent the frequencies of the data.

EXAMPLE 2–8 College Spending for First-Year Students

The table shows the average money spent by first-year college students. Draw a horizontal and vertical bar graph for the data.

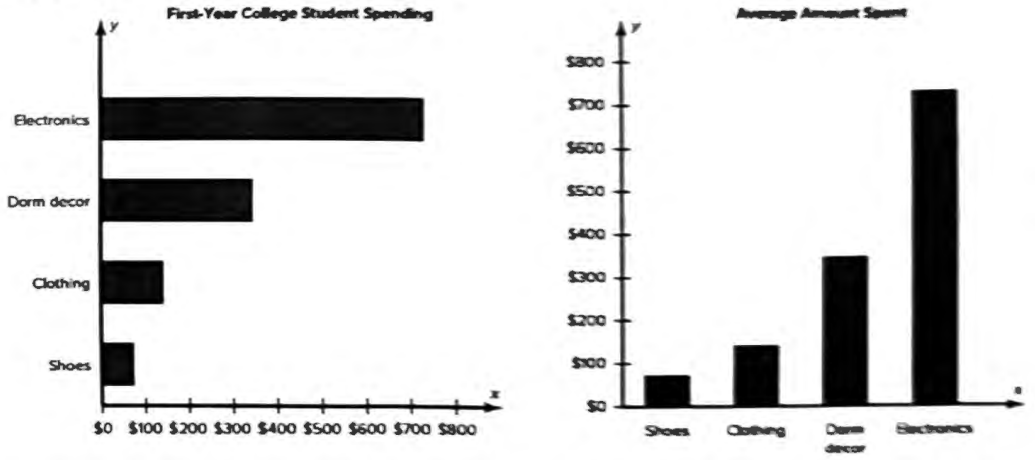
Electronics	$728
Dorm decor	344
Clothing	141
Shoes	72

Source: The National Retail Federation

SOLUTION

1. Draw and label the x and y axes. For the horizontal bar graph place the frequency scale on the x axis, and for the vertical bar graph place the frequency scale on the y axis.
2. Draw the bars corresponding to the frequencies. See Figure 2–9.

FIGURE 2–9 Bar Graphs for Example 2–8



The graphs show that first-year college students spend the most on electronic equipment.

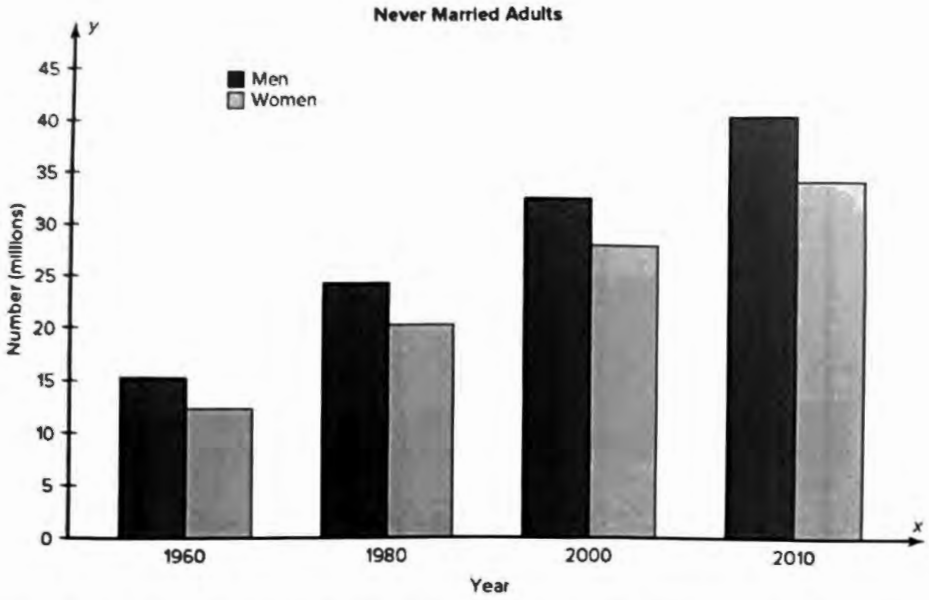
Bar graphs can also be used to compare data for two or more groups. These types of bar graphs are called *compound bar graphs*. Consider the following data for the number (in millions) of never married adults in the United States.

1960	15.3	12.3
1980	24.2	20.2
2000	32.3	27.8
2010	40.2	34.0

Source: U.S. Census Bureau.

Figure 2–10 shows a bar graph that compares the number of never married males with the number of never married females for the years shown. The comparison is made by placing the bars next to each other for the specific years. The heights of the bars can be compared. This graph shows that there have consistently been more never married

FIGURE 2–10
Example of a Compound Bar Graph



males than never married females and that the difference in the two groups has increased slightly over the last 50 years.

Pareto Charts

When the variable displayed on the horizontal axis is qualitative or categorical, a *Pareto chart* can also be used to represent the data.

A Pareto chart is used to represent a frequency distribution for a categorical variable, and the frequencies are displayed by the heights of vertical bars, which are arranged in order from highest to lowest.

EXAMPLE 2–9 Traffic Congestion

The data shown consist of the average number of hours that a commuter spends in traffic congestion per year in each city. Draw and analyze a Pareto chart for the data.

City	Hours
Atlanta	52
Boston	64
Chicago	61
New York	74
Washington, D. C.	82

Source: 2015 Urban Mobility Scorecard

SOLUTION

Step 1 Arrange the data from the largest to the smallest according to the number of hours.

City	Hours
Washington, D.C.	82
New York	74
Boston	64
Chicago	61
Atlanta	52

Historical Note

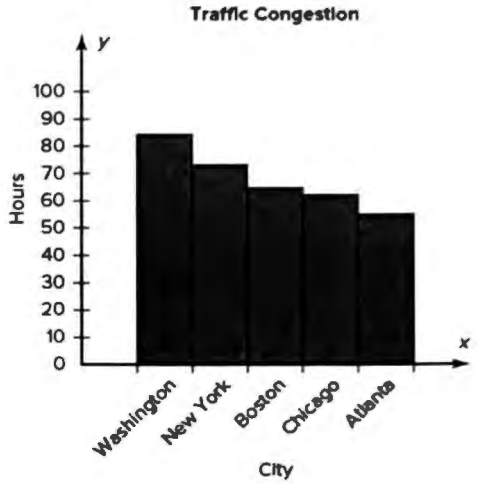
Vilfredo Pareto (1848–1923) was an Italian scholar who developed theories in economics, statistics, and the social sciences. His contributions to statistics include the development of a mathematical function used in economics. This function has many statistical applications and is called the Pareto distribution. In addition, he researched income distribution, and his findings became known as Pareto's law.

Step 2 Draw and label the x and y axes.

Step 3 Draw the vertical bars according to the number of hours (large to small).

The graph shown in Figure 2–11 shows that Washington, D.C. has the longest congestion times and Atlanta has the shortest times for the selection of cities.

FIGURE 2–11 Pareto Chart for Example 2–9



Suggestions for Drawing Pareto Charts

1. Make the bars the same width.
2. Arrange the data from largest to smallest according to frequency.
3. Make the units that are used for the frequency equal in size.

When you analyze a Pareto chart, make comparisons by looking at the heights of the bars.

The Time Series Graph

When data are collected over a period of time, they can be represented by a time series graph.

A time series graph represents data that occur over a specific period of time.

Example 2–10 shows the procedure for constructing a time series graph.

EXAMPLE 2–10 Price of an Advertisement for the Academy Awards Show

The data show the average cost (in millions of dollars) of a 30-second television ad on the Academy Awards show. Draw and analyze a time series graph for the data.

	2010	2011	2012	2013	2014	2015
	1.40	1.55	1.61	1.65	1.78	1.90

Source: Kantar Media, *USA TODAY* RESEARCH

Historical Note

Time series graphs are over 1000 years old. The first ones were used to chart the movements of the planets and the sun.

SOLUTION

Step 1 Draw and label the x and y axes.

Step 2 Label the x axis for years and label the y axis for cost.

Step 3 Plot each point for the values shown in the table.

Step 4 Draw line segments connecting adjacent points. Do not try to fit a smooth curve through the data points. See Figure 2–12.

The data show that there has been an increase every year. The largest increase (shown by the steepest line segment) occurred for the year 2011 compared to 2010. The increases for the years 2011, 2012, and 2013 were relatively small compared to the increases from 2010 to 2014 and 2014 to 2015.

FIGURE 2–12 Figure for Example 2–10

When you analyze a time series graph, look for a trend or pattern that occurs over the time period. For example, is the line ascending (indicating an increase over time) or descending (indicating a decrease over time)? Another thing to look for is the slope, or steepness, of the line. A line that is steep over a specific time period indicates a rapid increase or decrease over that period.

Two or more data sets can be compared on the same graph called a *compound time series graph* if two or more lines are used, as shown in Figure 2–13. This graph shows

FIGURE 2–13
Two Time Series Graphs for Comparison

Source: Bureau of Census, U.S. Department of Commerce.

the percentage of elderly males and females in the U.S. labor force from 1960 to 2010. It shows that the percentage of elderly men decreased significantly from 1960 to 1990 and then increased slightly after that. For the elderly females, the percentage decreased slightly from 1960 to 1980 and then increased from 1980 to 2010.

The Pie Graph

Pie graphs are used extensively in statistics. The purpose of the pie graph is to show the relationship of the parts to the whole by visually comparing the sizes of the sections. Percentages or proportions can be used. The variable is nominal or categorical.

A pie graph is a circle that is divided into sections or wedges according to the percentage of frequencies in each category of the distribution.

Example 2–11 shows the procedure for constructing a pie graph.

EXAMPLE 2–11 Super Bowl Snack Foods

This frequency distribution shows the number of pounds of each snack food eaten during the Super Bowl. Construct a pie graph for the data.

Snack	Pounds (frequency)
Potato chips	11.2 million
Tortilla chips	8.2 million
Pretzels	4.3 million
Popcorn	3.8 million
Snack nuts	2.5 million
	Total $n = 30.0$ million

Source: USA TODAY Weekend.

SOLUTION

Step 1 Since there are 360° in a circle, the frequency for each class must be converted to a proportional part of the circle. This conversion is done by using the formula

$$\text{Degrees} = \frac{f}{n} \cdot 360°$$

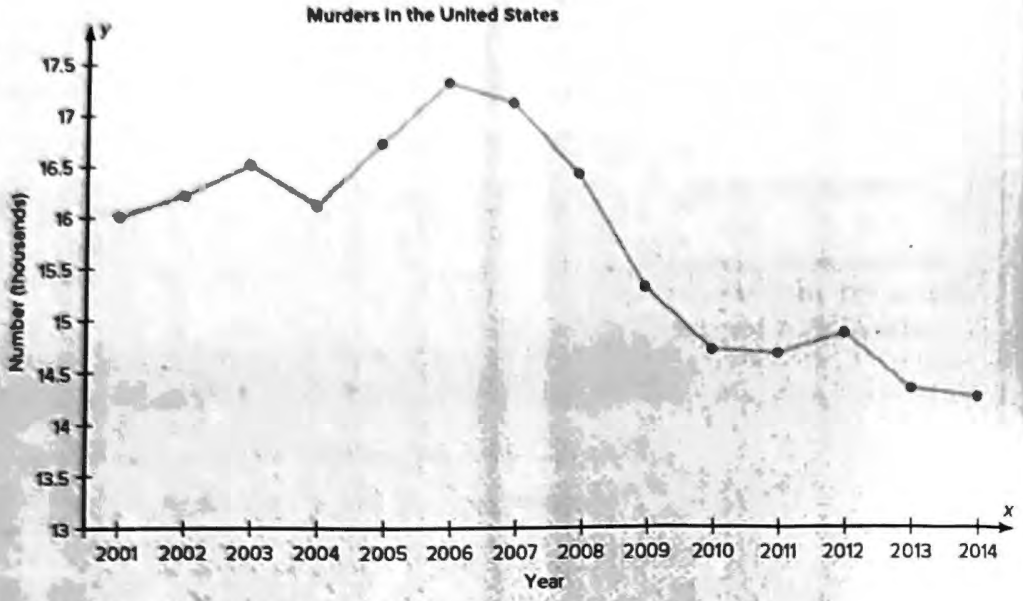
where f = frequency for each class and n = sum of the frequencies. Hence, the following conversions are obtained. The degrees should sum to 360°.[1]

Potato chips	$\frac{11.2}{30} \cdot 360° = 134°$
Tortilla chips	$\frac{8.2}{30} \cdot 360° = 98°$
Pretzels	$\frac{4.3}{30} \cdot 360° = 52°$
Popcorn	$\frac{3.8}{30} \cdot 360° = 46°$
Snack nuts	$\frac{2.5}{30} \cdot 360° = 30°$
Total	360°

[1]*Note:* The degrees column does not always sum to 360° due to rounding.

SPEAKING OF STATISTICS Murders in the United States

The graph shows the number of murders (in thousands) that have occurred in the United States since 2001. Based on the graph, do you think the number of murders is increasing, decreasing, or remaining the same?



Source: Crime in the United States 2015, FBI, Department of Justice.

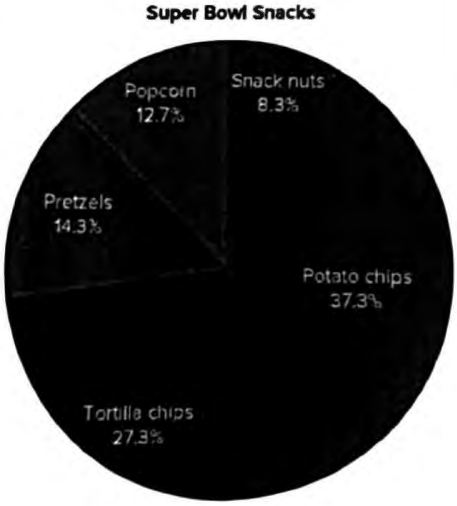
Step 2 Each frequency must also be converted to a percentage. Recall from Example 2–1 that this conversion is done by using the formula

$$\% = \frac{f}{n} \cdot 100$$

Hence, the following percentages are obtained. The percentages should sum to 100%.[2]

Potato chips	$\frac{11.2}{30} \cdot 100 = 37.3\%$	
Tortilla chips	$\frac{8.2}{30} \cdot 100 = 27.3\%$	
Pretzels	$\frac{4.3}{30} \cdot 100 = 14.3\%$	
Popcorn	$\frac{3.8}{30} \cdot 100 = 12.7\%$	
Snack nuts	$\frac{2.5}{30} \cdot 100 = 8.3\%$	
Total		99.9%

FIGURE 2–14 Pie Graph for Example 2–11



Step 3 Next, using a protractor and a compass, draw the graph, using the appropriate degree measures found in Step 1, and label each section with the name and percentages, as shown in Figure 2–14.

[2]*Note:* The percent column does not always sum to 100% due to rounding.

EXAMPLE 2–12 Police Calls

Construct and analyze a pie graph for the calls received each shift by a local municipality for a recent year. (Data obtained by author.)

Shift	Frequency
1. Day	2594
2. Evening	2800
3. Night	2436
	7830

SOLUTION

Step 1 Find the number of degrees for each shift, using the formula:

$$\text{Degrees} = \frac{f}{n} \cdot 360°$$

For each shift, the following results are obtained:

Day: $\frac{2594}{7830} \cdot 360° = 119°$

Evening: $\frac{2800}{7830} \cdot 360° = 129°$

Night: $\frac{2436}{7830} \cdot 360° = 112°$

Step 2 Find the percentages:

Day: $\frac{2594}{7830} \cdot 100 = 33\%$

Evening: $\frac{2800}{7830} \cdot 100 = 36\%$

Night: $\frac{2436}{7830} \cdot 100 = 31\%$

Step 3 Using a protractor, graph each section and write its name and corresponding percentage as shown in Figure 2–15.

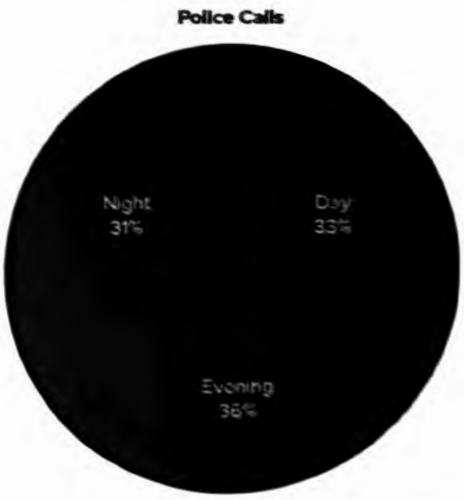


FIGURE 2–15
Figure for Example 2–12

To analyze the nature of the data shown in the pie graph, look at the size of the sections in the pie graph. For example, are any sections relatively large compared to the rest? Figure 2–15 shows that the number of calls for the three shifts are about equal, although slightly more calls were received on the evening shift.

Note: Computer programs can construct pie graphs easily, so the mathematics shown here would only be used if those programs were not available.

Dotplots

A dotplot uses points or dots to represent the data values. If the data values occur more than once, the corresponding points are plotted above one another.

A dotplot is a statistical graph in which each data value is plotted as a point (dot) above the horizontal axis.

Dotplots are used to show how the data values are distributed and to see if there are any extremely high or low data values.

EXAMPLE 2–13 Named Storms

The data show the number of named storms each year for the last 40 years. Construct and analyze a dotplot for the data.

19	15	14	7	6	11	11
9	16	8	8	11	9	8
16	12	13	14	13	12	7
15	15	19	11	4	6	13
10	15	7	12	6	10	
28	12	8	7	12	9	

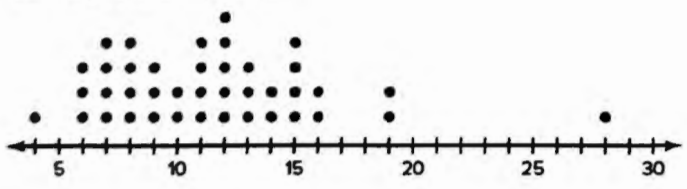
Source: NOAA.

Step 1 Find the lowest and highest data values, and decide what scale to use on the horizontal axis. The lowest data value is 4 and the highest data value is 28, so a scale from 4 to 28 is needed.

Step 2 Draw a horizontal line, and draw the scale on the line.

Step 3 Plot each data value above the line. If the value occurs more than once, plot the other point above the first point. See Figure 2–16.

FIGURE 2–16 Figure for Example 2–13



The graph shows that the majority of the named storms occur with frequency between 6 and 16 per year. There are only 3 years when there were 19 or more named storms per year.

Stem and Leaf Plots

The stem and leaf plot is a method of organizing data and is a combination of sorting and graphing. It has the advantage over a grouped frequency distribution of retaining the actual data while showing them in graphical form.

OBJECTIVE 4

Draw and interpret a stem and leaf plot.

A stem and leaf plot is a data plot that uses part of the data value as the stem and part of the data value as the leaf to form groups or classes.

For example, a data value of 34 would have 3 as the stem and 4 as the leaf. A data value of 356 would have 35 as the stem and 6 as the leaf.

Example 2–14 shows the procedure for constructing a stem and leaf plot.

EXAMPLE 2–14 Out Patient Cardiograms

At an outpatient testing center, the number of cardiograms performed each day for 20 days is shown. Construct a stem and leaf plot for the data.

25	31	20	32	13
14	43	02	57	23
36	32	33	32	44
32	52	44	51	45

SOLUTION

Step 1 Arrange the data in order:

02, 13, 14, 20, 23, 25, 31, 32, 32, 32,
32, 33, 36, 43, 44, 44, 45, 51, 52, 57

Note: Arranging the data in order is not essential and can be cumbersome when the data set is large; however, it is helpful in constructing a stem and leaf plot. The leaves in the final stem and leaf plot should be arranged in order.

Step 2 Separate the data according to the first digit, as shown.

02 13, 14 20, 23, 25 31, 32, 32, 32, 32, 33, 36
43, 44, 44, 45 51, 52, 57

Step 3 A display can be made by using the leading digit as the *stem* and the trailing digit as the *leaf*. For example, for the value 32, the leading digit, 3, is the stem and the trailing digit, 2, is the leaf. For the value 14, the 1 is the stem and the 4 is the leaf. Now a plot can be constructed as shown in Figure 2–17.

FIGURE 2–17
Stem and Leaf Plot for Example 2–14

Leading digit (stem)	Trailing digit (leaf)
0	2
1	3 4
2	0 3 5
3	1 2 2 2 2 3 6
4	3 4 4 5
5	1 2 7

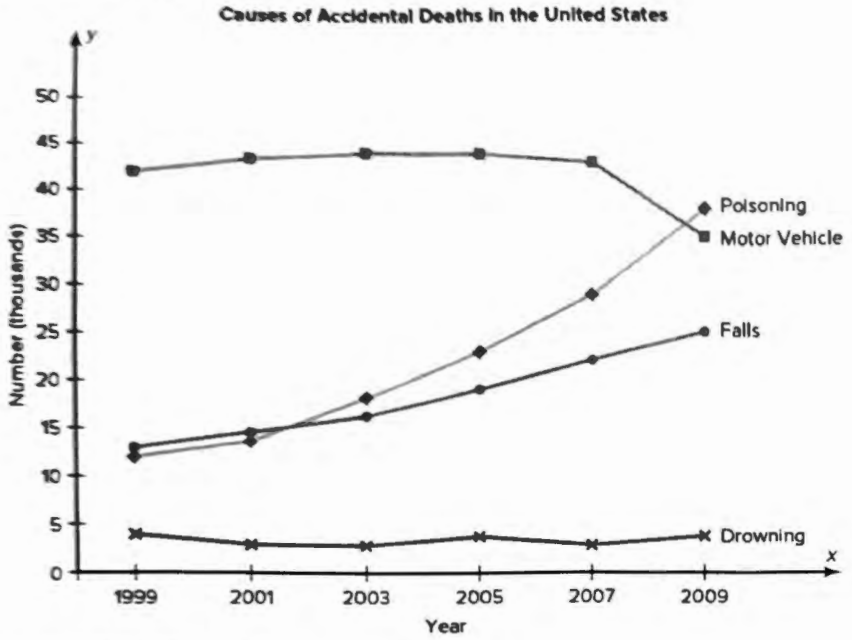
Figure 2–17 shows that the distribution peaks in the center and that there are no gaps in the data. For 7 of the 20 days, the number of patients receiving cardiograms was between 31 and 36. The plot also shows that the testing center treated from a minimum of 2 patients to a maximum of 57 patients in any one day.

If there are no data values in a class, you should write the stem number and leave the leaf row blank. Do not put a zero in the leaf row.

Applying the Concepts 2–3

Causes of Accidental Deaths in the United States, 1999–2009

The graph shows the number of deaths in the United States due to accidents. Answer the following questions about the graph.



Source: National Safety Council.

1. Name the variables used in the graph.
2. Are the variables qualitative or quantitative?
3. What type of graph is used here?
4. Which variable shows a decrease in the number of deaths over the years?
5. Which variable or variables show an increase in the number of deaths over the years?
6. The number of deaths in which variable remains about the same over the years?
7. List the approximate number of deaths for each category for the year 2001.
8. In 1999, which variable accounted for the most deaths? In 2009, which variable accounted for the most deaths?
9. In what year were the numbers of deaths from poisoning and falls about the same?

See page 108 for the answers.

Exercises 2–3

1. **Tech Company Employees** Construct a vertical and horizontal bar graph for the number of employees (in thousands) of a sample of the largest tech companies as of 2014.

Company	Employees
IBM	380
Hewlett Packard	302
Xerox	147
Microsoft	128
Intel	107

Source: S & P Capital IQ

2. **Worldwide Sales of Fast Foods** The worldwide sales (in billions of dollars) for several fast-food franchises for a specific year are shown. Construct a vertical bar graph and a horizontal bar graph for the data.

Wendy's	$ 8.7
KFC	14.2
Pizza Hut	9.3
Burger King	12.7
Subway	10.0

Source: Franchise Times.

3. **Gulf Coastlines** Construct a Pareto chart for the sizes of Gulf coastlines in statute miles for each state.

State	Coastline
Alabama	53
Florida	770
Louisiana	397
Mississippi	44
Texas	367

4. **Roller Coaster Mania** The World Roller Coaster Census Report lists the following numbers of roller coasters on each continent. Represent the data graphically, using a Pareto chart.

Africa	17
Asia	315
Australia	22
Europe	413
North America	643
South America	45

Source: www.rcdb.com

5. **Online Ad Spending** The amount spent (in billions of dollars) for ads online is shown. (The numbers for 2016 through 2019 are projected numbers.) Draw a time series graph and comment on the trend.

Year	2014	2015	2016	2017	2018	2019
Amount	$19.72	$31.53	$43.83	$53.29	$61.14	$69.04

Source: eMarketer.

6. **Violent Crimes** The number of all violent crimes (murder, nonnegligent homicide, manslaughter, forcible rape, robbery, and aggravated assault) in the United States for each of these years is listed below. Represent the data with a time series graph.

2000	1,425,486	2004	1,360,088	2008	1,394,461
2001	1,439,480	2005	1,390,745	2009	1,325,896
2002	1,423,677	2006	1,435,123	2010	1,246,248
2003	1,383,676	2007	1,422,970		

Source: World Almanac and Book of Facts.

7. **U.S. Licensed Drivers 70 or Older** Draw a time series graph for the number (in millions) of drivers in the United States 70 or older

Year	1982	1992	2002	2012
Number	10	15	20	23

Source: Federal Highway Administration

8. **Valentine's Day Spending** The data show the average amount of money spent by consumers on Valentine's Day. Draw a time series graph for the data and comment on the trend.

Year	2007	2008	2009	2010	2011	2012
Amount	$120	$123	$103	$103	$110	$126

Source: National Retail Federation.

9. **Credit Cards** Draw and analyze a pie graph for the number of credit cards a person has.

Number of cards	0	1	2 or 3	4 or more
Number	52	40	68	40

Source: Based on information from AARP Bulletin survey

10. **Reasons We Travel** The following data are based on a survey from American Travel Survey on why people travel. Construct a pie graph for the data and analyze the results.

Purpose	Number
Personal business	146
Visit friends or relatives	330
Work-related	225
Leisure	299

Source: USA TODAY

11. **Kids and Guns** The following data show where children obtain guns for committing crimes. Draw and analyze a pie graph for the data.

Source	Friend	Family	Street	Gun or Pawn Shop	Other
Number	24	15	9	9	6

12. **Colors of Automobiles** The popular car colors are shown. Construct a pie graph for the data.

White	19%
Silver	18
Black	16
Red	13
Blue	12
Gray	12
Other	10

Source: Dupont Automotive Color Popularity Report

13. **Ages of Football Players** The data show the ages of the players of the Super Bowl L Denver Bronco Champs in 2016. Construct a dotplot for the data, and comment on the distribution.

24	23	25	25	26	30
30	33	23	32	21	26
24	24	27	26	30	24
26	28	24	23	39	26
34	25	24	26	24	23
24	29	25	26	30	22
23	28	25	24	34	27
29	28	23	25	28	28
29	33	25	27	25	

Source: Funsided.com

14. **Teacher Strikes** In Pennsylvania the numbers of teacher strikes for the last 14 years are shown. Construct a dotplot for the data. Comment on the graph.

9	13	15	7	7	14	9
10	14	18	7	8	8	3

Source: School Leader News

15. Years of Experience The data show the number of years of experience the players on the Pittsburgh Steelers football team have at the beginning of the season. Draw and analyze a dot plot for the data.

4	4	2	9	7	3	7	12	6
5	1	4	5	2	7	6	12	3
12	4	0	4	0	0	0	2	9
2	6	7	13	4	2	6	9	4
4	0	3	5	4	2	6	9	4
4	0	3	5	3	11	1	4	2
3	15	1	6	0	11	3	10	3

16. Commuting Times Fifty off-campus students were asked how long it takes them to get to school. The times (in minutes) are shown. Construct a dotplot and analyze the data.

23	22	29	19	12
18	17	30	11	27
11	18	26	25	20
25	15	24	21	31
29	14	22	25	29
24	12	30	27	21
27	25	21	14	28
17	17	24	20	26
13	20	27	26	17
18	25	21	33	29

17. 50 Home Run Club There are 43 Major League baseball players (as of 2015) that have hit 50 or more home runs in one season. Construct a stem and leaf plot and analyze the data.

50	51	52	54	59	51	53
54	50	58	51	54	53	
56	58	56	70	54	52	
58	54	64	52	73	57	
50	60	56	50	66	54	
52	51	58	63	57	52	
51	50	61	52	65	50	

Source: The World Almanac and Book of Facts.

18. Calories in Salad Dressings A listing of calories per 1 ounce of selected salad dressings (not fat-free) is given below. Construct a stem and leaf plot for the data.

100	130	130	130	110	110	120	130	140	100
140	170	160	130	160	120	150	100	145	145
145	115	120	100	120	160	140	120	180	100
160	120	140	150	190	150	180	160		

19. Length of Major Rivers The data show the lengths (in hundreds of miles) of major rivers in South America and Europe. Construct a back-to-back stem and leaf plot, and compare the distributions.

South America				Europe				
39	21	10	10	5	12	7	6	8
11	10	2	10	5	5	4	6	
10	14	10	12	18	5	13	9	
17	15	10		14	6	6	11	
15	25	16		8	6	3	4	

Source: The World Almanac and Book of Facts.

20. Math and Reading Achievement Scores The math and reading achievement scores from the National Assessment of Educational Progress for selected states are listed below. Construct a back-to-back stem and leaf plot with the data, and compare the distributions.

Math					Reading				
52	66	69	62	61	65	76	76	66	67
63	57	59	59	55	71	70	70	66	61
55	59	74	72	73	61	69	78	76	77
68	76	73			77	77	80		

Source: World Almanac.

21. State which type of graph (Pareto chart, time series graph, or pie graph) would most appropriately represent the data.

a. Situations that distract automobile drivers
b. Number of persons in an automobile used for getting to and from work each day
c. Amount of money spent for textbooks and supplies for one semester
d. Number of people killed by tornados in the United States each year for the last 10 years
e. The number of pets (dogs, cats, birds, fish, etc.) in the United States this year
f. The average amount of money that a person spent for his or her significant other for Christmas for the last 6 years

22. State which graph (Pareto chart, time series graph, or pie graph) would most appropriately represent the given situation.

a. The number of students enrolled at a local college for each year during the last 5 years
b. The budget for the student activities department at a certain college for a specific year
c. The means of transportation the students use to get to school
d. The percentage of votes each of the four candidates received in the last election
e. The record temperatures of a city for the last 30 years
f. The frequency of each type of crime committed in a city during the year

23. U.S. Health Dollar The U.S. health dollar is spent as indicated below. Construct two different types of graphs to represent the data.

Government administration	9.7%
Nursing home care	5.5
Prescription drugs	10.1
Physician and clinical services	20.3
Hospital care	30.5
Other (OTC drugs, dental, etc.)	23.9

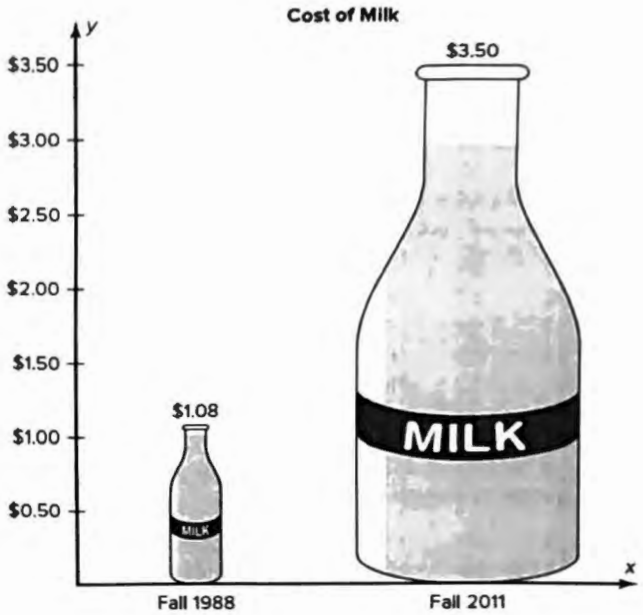
Source: Time Almanac.

24. Patents The U.S. Department of Commerce reports the following number of U.S. patents received by foreign countries and the United States in the year 2010. Illustrate the data with a bar graph and a pie graph. Which do you think better illustrates this data set?

Japan	44,814	United Kingdom	4,302
Germany	12,363	China	2,657
South Korea	11,671	Israel	1,819
Taiwan	8,238	Italy	1,796
Canada	4,852	United States	107,792

Source: World Almanac.

25. Cost of Milk The graph shows the increase in the price of a quart of milk. Why might the increase appear to be larger than it really is?



26. U.S. Population by Age The following information was found in a recent almanac. Use a pie graph to illustrate the information. Is there anything wrong with the data?

U.S. Population by Age in 2011

Under 20 years	27.0%
20 years and over	73.0
65 years and over	13.1

Source: Time Almanac.

27. Chicago Homicides Draw and compare two time series graphs for the number of homicides in the Chicago area.

Year	Homicides	As of June 29
2005	451	207
2007	448	204
2009	459	204
2011	435	187
2013	414	180

28. Trip Reimbursements The average amount requested for business trip reimbursement is itemized below. Illustrate the data with an appropriate graph. Do you have any questions regarding the data?

Flight	$440
Hotel stay	323
Entertainment	139
Phone usage	95
Transportation	65
Meal	38
Parking	34

Source: USA TODAY.

Technology Step by Step

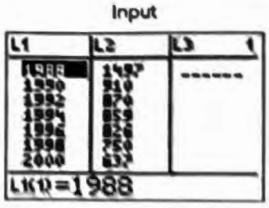
TI-84 Plus

Step by Step

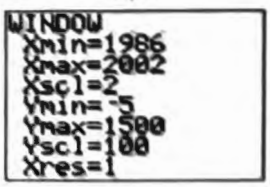
To graph a time series, follow the procedure for a frequency polygon from Section 2–2, using the following data for the number of outdoor drive-in theaters

	1988	1990	1992	1994	1996	1998	2000
	1497	910	870	859	826	750	637

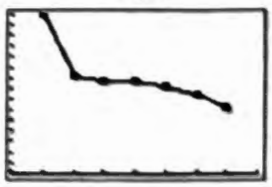
Input



Input



Output



Important Formulas

Formula for the percentage of values in each class:

$$\% = \frac{f}{n} \cdot 100$$

where

f = **frequency of class**

n = **total number of values**

Formula for the range:

R = **highest value − lowest value**

Formula for the class width:

Class width = upper boundary − lower boundary

Formula for the class midpoint:

$$X_m = \frac{\text{lower boundary + upper boundary}}{2}$$

or

$$X_m = \frac{\text{lower limit + upper limit}}{2}$$

Formula for the degrees for each section of a pie graph:

$$\text{Degrees} = \frac{f}{n} \cdot 360°$$

Review Exercises

Section 2–1

1. **How People Get Their News** The Brunswick Research Organization surveyed 50 randomly selected individuals and asked them the primary way they received the daily news. Their choices were via newspaper (N), television (T), radio (R), or Internet (I). Construct a categorical frequency distribution for the data and interpret the results.

N	N	T	T	T	I	R	R	I	T
I	N	R	R	I	N	N	I	T	N
I	R	T	T	T	T	N	R	R	I
R	R	I	N	T	R	T	I	I	T
T	I	N	T	T	I	R	N	R	T

2. **Men's World Hockey Champions** The United States won the Men's World Hockey Championship in 1933 and 1960. Below are listed the world champions for the last 30 years. Use this information to construct a frequency distribution of the champions. What is the difficulty with these data?

1982	USSR	1999	Czech Republic
1983	USSR	2000	Czech Republic
1984	Not held	2001	Czech Republic
1985	Czechoslovakia	2002	Slovakia
1986	USSR	2003	Canada
1987	Sweden	2004	Canada
1988	Not held	2005	Czech Republic
1989	USSR	2006	Sweden
1990	Sweden	2007	Canada
1991	Sweden	2008	Russia
1992	Sweden	2009	Russia
1993	Russia	2010	Czech Republic
1994	Canada	2011	Finland
1995	Finland	2012	Russia
1996	Czech Republic	2013	Sweden
1997	Canada	2014	Russia
1998	Sweden	2015	Canada

Source: Time Almanac

3. **BUN Count** The blood urea nitrogen (BUN) count of 20 randomly selected patients is given here in milligrams per deciliter (mg/dl). Construct an ungrouped frequency distribution for the data.

17	18	13	14
12	17	11	20
13	18	19	17
14	16	17	12
16	15	19	22

4. **Wind Speed** The data show the average wind speed for 36 days in a large city. Construct an ungrouped frequency distribution for the data.

8	15	9	8	9	10
8	10	14	9	8	8
12	9	8	8	14	9
9	13	13	10	12	9
13	8	11	11	9	8
9	13	9	8	8	10

5. **Waterfall Heights** The data show the heights (in feet) of notable waterfalls in North America. Organize the data into a grouped frequency distribution using 6 classes. This data will be used for Exercises 7, 9, and 11.

90	420	300	194
640	68	268	276
620	76	165	833
370	53	132	600
594	70	308	
574	215	109	
317	850	212	
300	256	187	

Source: National Geographic Society

6. **Ages of the Vice Presidents at the Time of Their Death** The ages at the time of death of those Vice Presidents of the United States who have passed away are listed below. Use the data to construct a frequency distribution.

Use 6 classes. The data for this exercise will be used for Exercises 8, 10, and 12.

90	83	80	73	70	51	68	79	70	71	
72	74	67	54	81	66	62	63	68	57	
66	96	78	55	60	66	57	71	60	85	
76	98	77	88	78	81	64	66	77	93	70

Source: World Almanac and Book of Facts.

Section 2–2

7. Find the relative frequency for the frequency distribution for the data in Exercise 5.

8. Find the relative frequency for the frequency distribution for the data in Exercise 6.

9. Construct a histogram, frequency polygon, and ogive for the data in Exercise 5.

10. Construct a histogram, frequency polygon, and ogive for the data in Exercise 6.

11. Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 5.

12. Construct a histogram, frequency polygon, and ogive, using relative frequencies for the data in Exercise 6.

Section 2–3

13. **Non-Alcoholic Beverages** The data show the yearly consumption (in gallons) of popular non-alcoholic beverages. Draw a vertical and horizontal bar graph to represent the data.

Soft drinks	52
Water	34
Milk	26
Coffee	21

Source: U.S. Department of Agriculture

14. **Calories of Nuts** The data show the number of calories per ounce in selected types of nuts. Construct vertical and horizontal bar graphs for the data.

Types	Calories
Peanuts	160
Almonds	170
Macadamia	200
Pecans	190
Cashews	160

15. **Crime** The data show the percentage of the types of crimes commonly committed in the United States. Construct a Pareto chart for the data.

Theft	55%
Burglary	20%
Motor Vehicle Theft	11%
Assault	8%
Rape & Homicide	1%

Source: FBI

16. **Pet Care** The data (in billions of dollars) show the estimated amount of money spent on pet care in the United States. Construct a Pareto chart for the data.

Type of care	Amount spent
Veterinarian care	$14
Supplies and medicine	11
Grooming and boarding	4
Animal purchases	2

Source: American Pet Products Association.

17. **Broadway Stage Engagements** The data show the number of new Broadway productions for the seasons. Construct and analyze a time series graph for the data.

Season	New Productions
2004	39
2005	39
2006	35
2007	39
2008	43
2009	39
2010	42
2011	41
2012	46
2013	44

Source: The Broadway League

18. **High School Dropout Rate** The data show the high school dropout rate for students for the years 2003 to 2013. Construct a time series graph and analyze the graph.

Year	Percent
2003	9.9
2004	10.3
2005	9.4
2006	9.3
2007	8.7
2008	8.0
2009	8.1
2010	7.4
2011	7.1
2012	6.6
2013	6.8

Source: U.S. Department of Commerce.

19. **Spending of College Freshmen** The average amounts spent by college freshmen for school items are shown. Construct a pie graph for the data.

Electronics/computers	$728
Dorm items	344
Clothing	141
Shoes	72

Source: National Retail Federation.

20. Smart Phone Insurance Construct and analyze a pie graph for the people who did or did not buy insurance for their smart phones at the time of purchase.

Response	Frequency
Yes	573
No	557
Don't Remember	166
Not offered any	211

Source: Based on information from Anderson Analytics

21. Peyton Manning's Colts Career Peyton Manning played for the Indianapolis Colts for 14 years. (He did not play in 2011.) The data show the number of touchdowns he scored for the years 1998–2010. Construct a dotplot for the data and comment on the graph.

26	33	27	49	31	27	33
26	26	29	28	31	33	

Source: NFL.com

22. Songs on CDs The data show the number of songs on each of 40 CDs from the author's collection. Construct a dotplot for the data and comment on the graph.

10	14	18	11
11	15	16	10
10	17	10	15
22	9	14	12
18	12	12	15
21	22	20	15
10	19	20	21
17	9	13	15
11	12	12	9
14	20	12	10

23. Weights of Football Players A local football team has 30 players; the weight of each player is shown. Construct a stem and leaf plot for the data. Use stems 20__, 21__, 22__, etc.

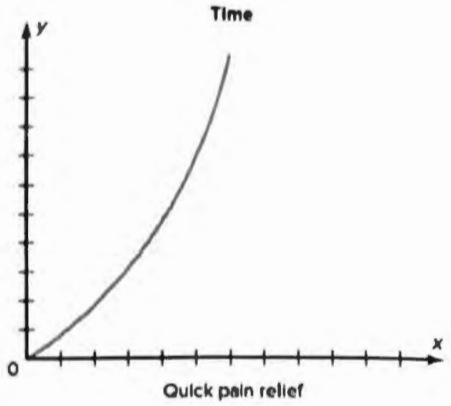
213	202	232	206	219
246	248	239	215	221
223	220	203	233	249
238	254	223	218	224
258	227	230	256	254
219	235	262	233	263

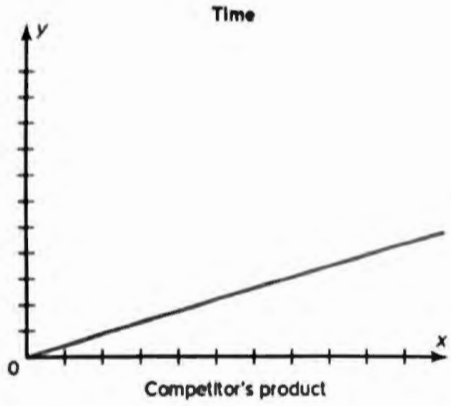
24. Public Libraries The numbers of public libraries in operation for selected states are listed below. Organize the data with a stem and leaf plot.

102	176	210	142	189	176	108	113	205
209	184	144	108	192	176			

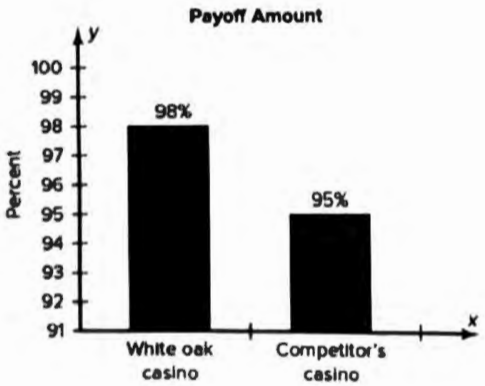
Source: World Almanac

25. Pain Relief The graph below shows the time it takes Quick Pain Relief to relieve a person's pain. The graph below that shows the time a competitor's product takes to relieve pain. Why might these graphs be misleading?





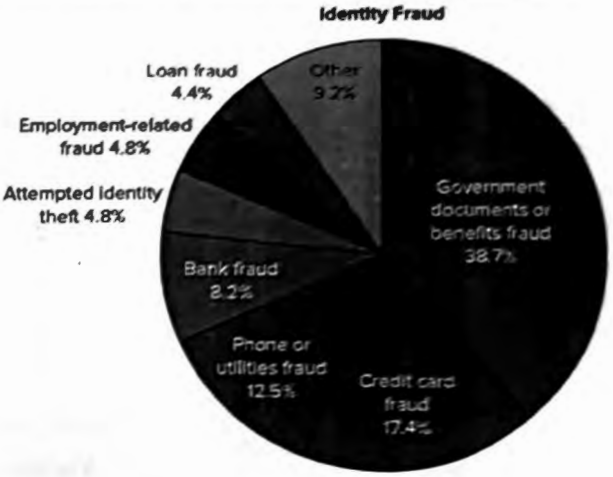
26. Casino Payoffs The graph shows the payoffs obtained from the White Oak Casino compared to the nearest competitor's casino. Why is this graph misleading?



STATISTICS TODAY

How Your Identity Can Be Stolen —Revisited

Data presented in numerical form do not convey an easy-to-interpret conclusion; however, when data are presented in graphical form, readers can see the visual impact of the numbers. In the case of identity fraud, the reader can see that most of the identity frauds are due to lost or stolen wallets, checkbooks, or credit cards, and very few identity frauds are caused by online purchases or transactions.



The Federal Trade Commission suggests some ways to protect your identity:

1. Shred all financial documents no longer needed.
2. Protect your Social Security number.
3. Don't give out personal information on the phone, through the mail, or over the Internet.
4. Never click on links sent in unsolicited emails.
5. Don't use an obvious password for your computer documents.
6. Keep your personal information in a secure place at home.

Data Analysis

A Data Bank is found in Appendix B, or on the World Wide Web by following links from www.mhhe.com/math/stat/bluman

1. From the Data Bank located in Appendix B, choose one of the following variables: age, weight, cholesterol level, systolic pressure, IQ, or sodium level. Select at least 30 values. For these values, construct a grouped frequency distribution. Draw a histogram, frequency polygon, and ogive for the distribution. Describe briefly the shape of the distribution.
2. From the Data Bank, choose one of the following variables: educational level, smoking status, or exercise. Select at least 20 values. Construct an ungrouped frequency distribution for the data. For the distribution, draw a Pareto chart and describe briefly the nature of the chart.
3. From the Data Bank, select at least 30 subjects and construct a categorical distribution for their marital status. Draw a pie graph and describe briefly the findings.
4. Using the data from Data Set IV in Appendix B, construct a frequency distribution and draw a histogram. Describe briefly the shape of the distribution of the tallest buildings in New York City.
5. Using the data from Data Set XI in Appendix B, construct a frequency distribution and draw a frequency polygon. Describe briefly the shape of the distribution for the number of pages in statistics books.

6. Using the data from Data Set IX in Appendix B, divide the United States into four regions, as follows:

Northeast	CT ME MA NH NJ NY PA RI VT
Midwest	IL IN IA KS MI MN MD MS NE ND OH SD WI
South	AL AR DE DC FL GA KY LA MD NC OK SC TN TX VA WV
West	AK AZ CA CO HI ID MT NV NM OR UT WA WY

Find the total population for each region, and draw a Pareto chart and a pie graph for the data. Analyze the results. Explain which chart might be a better representation for the data.

7. Using the data from Data Set I in Appendix B, make a stem and leaf plot for the record low temperatures in the United States. Describe the nature of the plot.

Chapter Quiz

Determine whether each statement is true or false. If the statement is false, explain why.

1. In the construction of a frequency distribution, it is a good idea to have overlapping class limits, such as 10–20, 20–30, 30–40.
2. Bar graphs can be drawn by using vertical or horizontal bars.
3. It is not important to keep the width of each class the same in a frequency distribution.
4. Frequency distributions can aid the researcher in drawing charts and graphs.
5. The type of graph used to represent data is determined by the type of data collected and by the researcher's purpose.
6. In construction of a frequency polygon, the class limits are used for the *x* axis.
7. Data collected over a period of time can be graphed by using a pie graph.

Select the best answer.

8. What is another name for the ogive?
   a. Histogram
   b. Frequency polygon
   c. Cumulative frequency graph
   d. Pareto chart
9. What are the boundaries for 8.6–8.8?
   a. 8–9
   b. 8.5–8.9
   c. 8.55–8.85
   d. 8.65–8.75
10. What graph should be used to show the relationship between the parts and the whole?
    a. Histogram
    b. Pie graph
    c. Pareto chart
    d. Ogive
11. Except for rounding errors, relative frequencies should add up to what sum?
    a. 0
    b. 1
    c. 50
    d. 100

Complete these statements with the best answers.

12. The three types of frequency distributions are _____, _____, and _____.
13. In a frequency distribution, the number of classes should be between _____ and _____.
14. Data such as blood types (A, B, AB, O) can be organized into a(n) _____ frequency distribution.
15. Data collected over a period of time can be graphed using a(n) _____ graph.
16. A statistical device used in exploratory data analysis that is a combination of a frequency distribution and a histogram is called a(n) _____.
17. On a Pareto chart, the frequencies should be represented on the _____ axis.
18. **Housing Arrangements** A questionnaire on housing arrangements showed this information obtained from 25 respondents. Construct a frequency distribution for the data (H = house, A = apartment, M = mobile home, C = condominium). These data will be used in Exercise 19.

H	C	H	M	H	A	C	A	M
C	M	C	A	M	A	C	C	M
C	C	H	A	H	H	M		

19. Construct a pie graph for the data in Exercise 18.
20. **Items Purchased at a Convenience Store** When 30 randomly selected customers left a convenience store, each was asked the number of items he or she purchased. Construct an ungrouped frequency

distribution for the data. These data will be used in Exercise 21.

2	9	4	3	6
6	2	8	6	5
7	5	3	8	6
6	2	3	2	4
6	9	9	8	9
4	2	1	7	4

21. Construct a histogram, a frequency polygon, and an ogive for the data in Exercise 20.

22. Coal Consumption The following data represent the energy consumption of coal (in billions of Btu) by each of the 50 states and the District of Columbia. Use the data to construct a frequency distribution and a relative frequency distribution with 7 classes.

631	723	267	60	372	15	19	92	306	38
413	8	736	156	478	264	1015	329	679	1498
52	1365	142	423	365	350	445	776	1267	0
26	356	173	373	335	34	937	250	33	84
0	253	84	1224	743	582	2	33	0	426
474									

Source: Time Almanac.

23. Construct a histogram, frequency polygon, and ogive for the data in Exercise 22. Analyze the histogram.

24. Recycled Trash Construct a Pareto chart and a horizontal bar graph for the number of tons (in millions) of trash recycled per year by Americans based on an Environmental Protection Agency study.

Type	Amount
Paper	320.0
Iron/steel	292.0
Aluminum	276.0
Yard waste	242.4
Glass	196.0
Plastics	41.6

Source: USA TODAY.

25. Identity Thefts The results of a survey of 84 people whose identities were stolen using various methods are shown. Draw a pie chart for the information.

Lost or stolen wallet, checkbook, or credit card	38
Retail purchases or telephone transactions	15
Stolen mail	9
Computer viruses or hackers	8
Phishing	4
Other	10
	84

Source: Javelin Strategy and Research.

26. Needless Deaths of Children *The New England Journal of Medicine* predicted the number of needless deaths due to childhood obesity. Draw a time series graph for the data.

Year	2020	2025	2030	2035
Deaths	130	550	1500	3700

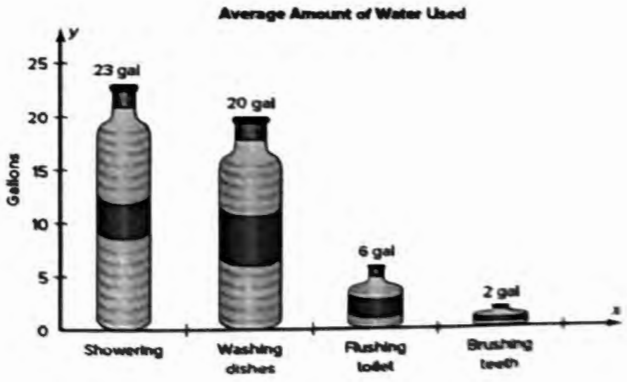
27. Museum Visitors The number of visitors to the Historic Museum for 25 randomly selected hours is shown. Construct a stem and leaf plot for the data.

15	53	48	19	38
86	63	98	79	38
62	89	67	39	26
28	35	54	88	76
31	47	53	41	68

28. Parking Meter Revenue In a small city the number of quarters collected from the parking meters is shown. Construct a dotplot for the data.

13	12	11	7	16
10	16	15	7	11
3	5	14	3	6
8	3	10	9	3
5	7	8	9	9
9	2	6	4	11
7	4	2	8	10
7	17	4	11	8
2	5	5	14	6
3	9	3	12	3

29. Water Usage The graph shows the average number of gallons of water a person uses for various activities. Can you see anything misleading about the way the graph is drawn?



Critical Thinking Challenges

1. The Great Lakes Shown are various statistics about the Great Lakes. Using appropriate graphs (your choice) and summary statements, write a report analyzing the data.

	Superior	Michigan	Huron	Erie	Ontario
Length (miles)	350	307	206	241	193
Breadth (miles)	160	118	183	57	53
Depth (feet)	1,330	923	750	210	802
Volume (cubic miles)	2,900	1,180	850	116	393
Area (square miles)	31,700	22,300	23,000	9,910	7,550
Shoreline (U.S., miles)	863	1,400	580	431	300

Source: The World Almanac and Book of Facts

2. Teacher Strikes In Pennsylvania there were more teacher strikes in 2004 than there were in all other states combined. Because of the disruptions, state legislators want to pass a bill outlawing teacher strikes and submitting contract disputes to binding arbitration. The graph shows the number of teacher strikes in Pennsylvania for the school years 1997 to 2011. Use the graph to answer these questions.

a. In what year did the largest number of strikes occur? How many were there?

b. In what year did the smallest number of teacher strikes occur? How many were there?

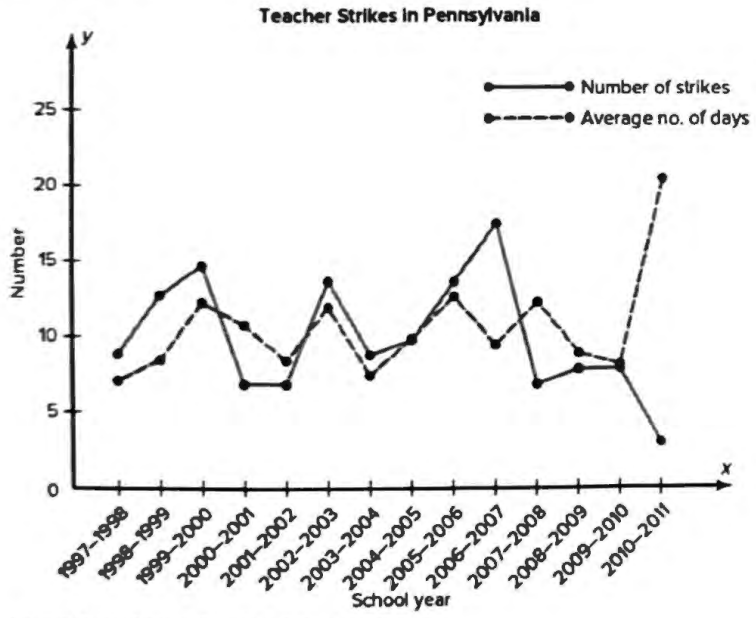
c. In what year was the average duration of the strikes the longest? How long was it?

d. In what year was the average duration of the strikes the shortest? How long was it?

e. In what year was the number of teacher strikes the same as the average duration of the strikes?

f. Find the difference in the number of strikes for the school years 1997–1998 and 2010–2011.

g. Do you think teacher strikes should be outlawed? Justify your conclusions.



Source: Pennsylvania School Boards Association.

Data Projects

Where appropriate, use the TI-84 Plus, Excel, MINITAB, or a computer program of your choice to complete the following exercises.

1. Business and Finance Consider the 30 stocks listed as the Dow Jones Industrials. For each, find its earnings per share. Randomly select 30 stocks traded on the NASDAQ. For each, find its earnings per share. Create a frequency table with 5 categories for each data set. Sketch a histogram for each. How do the two data sets compare?

3

Data Description

Alan Schein/Alamy RF

STATISTICS TODAY

How Long Are You Delayed by Road Congestion?

No matter where you live, at one time or another, you have been stuck in traffic. To see whether there are more traffic delays in some cities than in others, statisticians make comparisons using descriptive statistics. A statistical study by the Texas Transportation Institute found that a driver is delayed by road congestion an average of 36 hours per year. To see how selected cities compare to this average, see Statistics Today—Revisited at the end of the chapter.

This chapter will show you how to obtain and interpret descriptive statistics such as measures of average, measures of variation, and measures of position.

OUTLINE

Introduction

3–1 Measures of Central Tendency

3–2 Measures of Variation

3–3 Measures of Position

3–4 Exploratory Data Analysis

Summary

OBJECTIVES

After completing this chapter, you should be able to:

1. Summarize data, using measures of central tendency, such as the mean, median, mode, and midrange.
2. Describe data, using measures of variation, such as the range, variance, and standard deviation.
3. Identify the position of a data value in a data set, using various measures of position, such as percentiles, deciles, and quartiles.
4. Use the techniques of exploratory data analysis, including boxplots and five-number summaries, to discover various aspects of data.

Introduction

Chapter 2 showed how you can gain useful information from raw data by organizing them into a frequency distribution and then presenting the data by using various graphs. This chapter shows the statistical methods that can be used to summarize data. The most familiar of these methods is the finding of averages.

For example, you may read that the average speed of a car crossing midtown Manhattan during the day is 5.3 miles per hour or that the average number of minutes an American father of a 4-year-old spends alone with his child each day is 42.[1]

In the book *American Averages* by Mike Feinsilber and William B. Meed, the authors state:

> *"Average" when you stop to think of it is a funny concept. Although it describes all of us it describes none of us. . . . While none of us wants to be the average American, we all want to know about him or her.*

Interesting Fact

A person has an average of 1460 dreams in one year

The authors go on to give examples of averages:

> *The average American man is five feet, nine inches tall; the average woman is five feet, 3.6 inches.*
> *The average American is sick in bed seven days a year missing five days of work.*
> *On the average day, 24 million people receive animal bites.*
> *By his or her 70th birthday, the average American will have eaten 14 steers, 1050 chickens, 3.5 lambs, and 25.2 hogs.*[2]

In these examples, the word *average* is ambiguous, since several different methods can be used to obtain an average. Loosely stated, the average means the center of the distribution or the most typical case. Measures of average are also called *measures of central tendency* and include the *mean, median, mode,* and *midrange.*

Knowing the average of a data set is not enough to describe the data set entirely. Even though a shoe store owner knows that the average size of a man's shoe is size 10, she would not be in business very long if she ordered only size 10 shoes.

© Image/Getty Images RF

[1] "Harper's Index." *Harper's magazine*

[2] Mike Feinsilber and William B. Meed, *American Averages* (New York: Bantam Doubleday Dell).

As this example shows, in addition to knowing the average, you must know how the data values are dispersed. That is, do the data values cluster around the average, or are they spread more evenly throughout the distribution? The measures that determine the spread of the data values are called *measures of variation*, or *measures of dispersion*. These measures include the *range*, *variance*, and *standard deviation*.

Finally, another set of measures is necessary to describe data. These measures are called *measures of position*. They tell where a specific data value falls within the data set or its relative position in comparison with other data values. The most common position measures are *percentiles*, *deciles*, and *quartiles*. These measures are used extensively in psychology and education. Sometimes they are referred to as *norms*.

The measures of central tendency, variation, and position explained in this chapter are part of what is called *traditional statistics*.

Section 3–4 shows the techniques of what is called *exploratory data analysis*. These techniques include the *boxplot* and the *five-number summary*. They can be used to explore data to see what they show (as opposed to the traditional techniques, which are used to confirm conjectures about the data).

3–1 Measures of Central Tendency

OBJECTIVE 1

Summarize data, using measures of central tendency, such as the mean, median, mode, and midrange.

Chapter 1 stated that statisticians use samples taken from populations; however, when populations are small, it is not necessary to use samples since the entire population can be used to gain information. For example, suppose an insurance manager wanted to know the average weekly sales of all the company's representatives. If the company employed a large number of salespeople, say, nationwide, he would have to use a sample and make an inference to the entire sales force. But if the company had only a few salespeople, say, only 87 agents, he would be able to use all representatives' sales for a randomly chosen week and thus use the entire population.

Measures found by using all the data values in the population are called *parameters*. Measures obtained by using the data values from samples are called *statistics;* hence, the average of the sales from a sample of representatives is a *statistic*, and the average of sales obtained from the entire population is a *parameter*.

Historical Note

In 1796, Adolphe Quetelet investigated the characteristics (heights, weights, etc.) of French conscripts to determine the "average man." Florence Nightingale was so influenced by Quetelet's work that she began collecting and analyzing medical records in the military hospitals during the Crimean War. Based on her work, hospitals began keeping accurate records on their patients.

A statistic is a characteristic or measure obtained by using the data values from a sample.

A parameter is a characteristic or measure obtained by using all the data values from a specific population.

These concepts as well as the symbols used to represent them will be explained in detail in this chapter.

General Rounding Rule In statistics the basic rounding rule is that when computations are done in the calculation, rounding should not be done until the final answer is calculated. When rounding is done in the intermediate steps, it tends to increase the difference between that answer and the exact one. But in the textbook and solutions manual, it is not practical to show long decimals in the intermediate calculations; hence, the values in the examples are carried out to enough places (usually three or four) to obtain the same answer that a calculator would give after rounding on the last step.

There are specific rounding rules for many statistics, and they will be given in the appropriate sections.

The Mean

The *mean*, also known as the *arithmetic average*, is found by adding the values of the data and dividing by the total number of values. For example, the mean of 3, 2, 6, 5, and 4 is found by adding $3 + 2 + 6 + 5 + 4 = 20$ and dividing by 5; hence, the mean of the data is $20 \div 5 = 4$. The values of the data are represented by X's. In this data set, $X_1 = 3$, $X_2 = 2$, $X_3 = 6$,

$X_4 = 5$, and $X_5 = 4$. To show a sum of the total X values, the symbol Σ (the capital Greek letter sigma) is used, and ΣX means to find the sum of the X values in the data set. The summation notation is explained in the online resource section under "Algebra Review."

The **mean** is the sum of the values, divided by the total number of values.

The **sample mean,** denoted by $\bar{X}$ (pronounced "X bar"), is calculated by using sample data. The sample mean is a statistic.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\Sigma X}{n}$$

where n represents the total number of values in the sample.

The **population mean,** denoted by μ (pronounced "mew"), is calculated by using all the values in the population. The population mean is a parameter.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\Sigma X}{N}$$

where N represents the total number of values in the population.

In statistics, Greek letters are used to denote parameters, and Roman letters are used to denote statistics. Assume that the data are obtained from samples unless otherwise specified.

Rounding Rule for the Mean The mean should be rounded to one more decimal place than occurs in the raw data. For example, if the raw data are given in whole numbers, the mean should be rounded to the nearest tenth. If the data are given in tenths, the mean should be rounded to the nearest hundredth, and so on.

EXAMPLE 3–1 Avian Flu Cases

The number of confirmed flu cases for a 9-year period is shown. Find the mean.

4 46 98 115 88 44 73 48 62

Source: World Health Organization.

SOLUTION

$$\bar{X} = \frac{\Sigma X}{n} = \frac{4 + 46 + 98 + 115 + 88 + 44 + 73 + 48 + 62}{n}$$
$$= \frac{578}{9} \approx 64.2$$

Hence, the mean number of flu cases over the 9-year period is 64.2.

EXAMPLE 3–2 Store Sales

The data show the systemwide sales (in millions) for U.S. franchises of a well-known donut store for a 5-year period. Find the mean.

$221 $239 $262 $281 $318

Source: Krispy Kreme.

SOLUTION

$$\bar{X} = \frac{\Sigma X}{n} = \frac{221 + 239 + 262 + 281 + 318}{5} = \frac{1321}{5} = 264.2$$

The mean amount of sales for the stores over the 5-year period is $264.2 million.

The mean, in most cases, is not an actual data value.

The procedure for finding the mean for grouped data assumes that the mean of all the raw data values in each class is equal to the midpoint of the class. In reality, this is not true,

since the average of the raw data values in each class usually will not be exactly equal to the midpoint. However, using this procedure will give an acceptable approximation of the mean, since some values fall above the midpoint and other values fall below the midpoint for each class, and the midpoint represents an estimate of all values in the class.

The steps for finding the mean for grouped data are shown in the next Procedure Table.

Procedure Table

Finding the Mean for Grouped Data

Step 1 Make a table as shown.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$

Step 2 Find the midpoints of each class and place them in column C.

Step 3 Multiply the frequency by the midpoint for each class, and place the product in column D.

Step 4 Find the sum of column D.

Step 5 Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n}$$

[*Note:* The symbols $\Sigma f \cdot X_m$ mean to find the sum of the product of the frequency (f) and the midpoint (X_m) for each class.]

EXAMPLE 3–3 Salaries of CEOs

The frequency distribution shows the salaries (in millions) for a specific year of the top 25 CEOs in the United States. Find the mean.

Source: S & P Capital.

Class boundaries	Frequency
15.5–20.5	13
20.5–25.5	6
25.5–30.5	4
30.5–35.5	1
35.5–40.5	1
	Total 25

SOLUTION

15.5–20.5	13		
20.5–25.5	6		
25.5–30.5	4		
30.5–35.5	1		
35.5–40.5	1		
	$n = 25$		

Interesting Fact

The average time it takes a person to find a new job is 5.9 months.

SPEAKING OF STATISTICS — Ages of the Top 50 Wealthiest People

The histogram shows the ages of the top 50 wealthiest individuals according to *Forbes Magazine* for a recent year. The mean age is 66 years. The median age is 68 years. Explain why these two statistics are not enough to adequately describe the data.

© Don Farrall/Getty Images RF

Unusual Stat

A person looks, on average, at about 14 homes before he or she buys one.

$$X_m = \frac{15.5 + 20.5}{2} = 18 \qquad \frac{20.5 + 25.5}{2} = 23 \quad \text{etc.}$$

$$13 \cdot 18 = 234 \qquad 6 \cdot 23 = 138 \quad \text{etc.}$$

A Class	B Frequency	C Midpoint X_m	D $f \cdot X_m$
15.5–20.5	13	18	234
20.5–25.5	6	23	138
25.5–30.5	4	28	112
30.5–35.5	1	33	33
35.5–40.5	1	38	38
	$n = 25$		$\sum f \cdot X_m = 555$

$$\bar{X} = \frac{\sum f \cdot X_m}{n} = \frac{555}{25} = \$22.2 \text{ million}$$

The mean salary is $22.2 million.

Historical Note

The concept of median was used by Gauss at the beginning of the 19th century and introduced as a statistical concept by Francis Galton around 1874. The mode was first used by Karl Pearson in 1894.

The Median

An article recently reported that the median income for college professors was $43,250. This measure of central tendency means that one-half of all the professors surveyed earned more than $43,250, and one-half earned less than $43,250.

The *median* is the halfway point in a data set. Before you can find this point, the data must be arranged in ascending or increasing order. When the data set is ordered, it is called a **data array.** The median either will be a specific value in the data set or will fall between two values, as shown in the next examples.

The **median** is the midpoint of the data array. The symbol for the median is MD.

The Procedure Table for finding the median is shown next.

Procedure Table

Finding the Median

Step 1 Arrange the data values in ascending order.

Step 2 Determine the number of values in the data set.

Step 3 *a.* If n is odd, select the middle data value as the median.

b. If n is even, find the mean of the two middle values. That is, add them and divide the sum by 2.

EXAMPLE 3–4 Tablet Sales

The data show the number of tablet sales in millions of units for a 5-year period. Find the median of the data.

108.2 17.6 159.8 69.8 222.6

Source: Gartner.

SOLUTION

Step 1 Arrange the data values in ascending order.

17.6 69.8 108.2 159.8 222.6

Step 2 There are an odd number of data values, namely, 5.

Step 3 Select the middle data value.

17.6 69.8 108.2 159.8 222.6

↑

Median

The median of the number of tablet sales for the 5-year period is 108.2 million.

EXAMPLE 3–5 Tornadoes in the United States

The number of tornadoes that have occurred in the United States over an 8-year period is as follows. Find the median.

684, 764, 656, 702, 856, 1133, 1132, 1303

Source: The Universal Almanac.

SOLUTION

Step 1 Arrange the data values in ascending order.

656, 684, 702, 764, 856, 1132, 1133, 1303

Step 2 There are an even number of data values, namely, 8.

Step 3 The middle two data values are 764 and 856.

656, 684, 702, 764, 856, 1132, 1133, 1303

↑

Median

Since the middle point falls halfway between 764 and 856, find the median MD by adding the two values and dividing by 2.

$$\text{MD} = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

The Mode

The third measure of average is called the *mode*. The mode is the value that occurs most often in the data set. It is sometimes said to be the most typical case.

The value that occurs most often in a data set is called the mode.

A data set that has only one value that occurs with the greatest frequency is said to be **unimodal.**

If a data set has two values that occur with the same greatest frequency, both values are considered to be the mode and the data set is said to be **bimodal.** If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be **multimodal.** When no data value occurs more than once, the data set is said to have *no mode. Note: Do not say that the mode is zero.* That would be incorrect, because in some data, such as temperature, zero can be an actual value. A data set can have more than one mode or no mode at all. These situations will be shown in some of the examples that follow.

EXAMPLE 3–6 Public Libraries

The data show the number of public libraries in a sample of eight states. Find the mode.

114 77 21 101 311 77 159 382

Source: The World Almanac.

SOLUTION

It is helpful to arrange the data in order, although it is not necessary.

21 77 77 101 114 159 311 382

Since 77 occurs twice, a frequency larger than that of any other number, the mode is 77.

EXAMPLE 3–7 Licensed Nuclear Reactors

The data show the number of licensed nuclear reactors in the United States for a recent 15-year period. Find the mode.

Source: The World Almanac and Book of Facts.

104	104	104	104	104
107	109	109	109	110
109	111	112	111	109

SOLUTION

Since the values 104 and 109 both occur 5 times, the modes are 104 and 109. The data set is said to be bimodal.

EXAMPLE 3–8 U.S. Patent Leaders

The data show the number of patents secured for the top 5 companies for a specific year. Find the mode.

6180 4894 2821 2559 2483

Source: IFI Claims Patent Services.

Since each value occurs only once, there is no mode.

The mode for grouped data is the modal class. The **modal class** is the class with the largest frequency.

EXAMPLE 3–9 Salaries of CEOs

Find the modal class for the frequency distribution for the salaries of the top CEOs in the United States, shown in Example 3–3.

SOLUTION

Class	Frequency
15.5–20.5	13 ←Modal class
20.5–25.5	6
25.5–30.5	4
30.5–35.5	1
35.5–40.5	1

Since the class 15.5–20.5 has the largest frequency, 13, it is the modal class. Sometimes the midpoint of the class is used. In this case, it is 18.

The mode is the only measure of central tendency that can be used in finding the most typical case when the data are nominal or categorical.

EXAMPLE 3–10 Nonalcoholic Beverages

The data show the number of gallons of various nonalcoholic drinks Americans consume in a year. Find the mode.

Drink	Gallons
Soft drinks	52
Water	34
Milk	26
Coffee	21

Source: U.S. Department of Agriculture.

SOLUTION

Since the category of soft drinks has the largest frequency, 52, we can say that the mode or most typical drink is a soft drink.

An extremely high or extremely low data value in a data set can have a striking effect on the mean of the data set. These extreme values are called *outliers*. This is one reason why, when analyzing a frequency distribution, you should be aware of any of these values. For the data set shown in Example 3–11, the mean, median, and mode can be quite different because of extreme values. A method for identifying outliers is given in Section 3–3.

EXAMPLE 3–11 Salaries of Personnel

A small company consists of the owner, the manager, the salesperson, and two technicians, all of whose annual salaries are listed here. (Assume that this is the entire population.)

Staff	Salary
Owner	$100,000
Manager	40,000
Salesperson	24,000
Technician	18,000
Technician	18,000

Find the mean, median, and mode.

SOLUTION

$$\mu = \frac{\Sigma X}{N} = \frac{\$100{,}000 + 40{,}000 + 24{,}000 + 18{,}000 + 18{,}000}{5} = \frac{\$200{,}000}{5} = \$40{,}000$$

Hence, the mean is $40,000, the median is $24,000, and the mode is $18,000.

In Example 3–11, the mean is much higher than the median or the mode. This is so because the extremely high salary of the owner tends to raise the value of the mean. In this and similar situations, the median should be used as the measure of central tendency.

The Midrange

The *midrange* is a rough estimate of the middle. It is found by adding the lowest and highest values in the data set and dividing by 2. It is a very rough estimate of the average and can be affected by one extremely high or low value.

The midrange is defined as the sum of the lowest and highest values in the data set, divided by 2. The symbol MR is used for the midrange.

$$\text{MR} = \frac{\text{lowest value} + \text{highest value}}{2}$$

EXAMPLE 3–12 Bank Failures

The number of bank failures for a recent five-year period is shown. Find the midrange.

3, 30, 148, 157, 71

Source: Federal Deposit Insurance Corporation.

SOLUTION

The lowest data value is 3, and the highest data value is 157.

$$\text{MR} = \frac{3 + 157}{2} = \frac{160}{2} = 80$$

The midrange for the number of bank failures is 80.

EXAMPLE 3–13 **NFL Signing Bonuses**

Find the midrange of data for the NFL signing bonuses. The bonuses in millions of dollars are

18, 14, 34.5, 10, 11.3, 10, 12.4, 10

SOLUTION

The lowest bonus is \$10 million, and the largest bonus is \$34.5 million.

$$\text{MR} = \frac{10 + 34.5}{2} = \frac{44.5}{2} = \$22.25 \text{ million}$$

Notice that this amount is larger than seven of the eight amounts and is not typical of the average of the bonuses. The reason is that there is one very high bonus, namely, \$34.5 million.

In statistics, several measures can be used for an average. The most common measures are the mean, median, mode, and midrange. Each has its own specific purpose and use. Exercises 36 through 38 show examples of other averages, such as the harmonic mean, the geometric mean, and the quadratic mean. Their applications are limited to specific areas, as shown in the exercises.

The Weighted Mean

Sometimes, you must find the mean of a data set in which not all values are equally represented. Consider the case of finding the average cost of a gallon of gasoline for three taxis. Suppose the drivers buy gasoline at three different service stations at a cost of \$3.22, \$3.53, and \$3.63 per gallon. You might try to find the average by using the formula

$$\bar{X} = \frac{\sum X}{n}$$

$$= \frac{3.22 + 3.53 + 3.63}{3} = \frac{10.38}{3} = \$3.46$$

But not all drivers purchased the same number of gallons. Hence, to find the true average cost per gallon, you must take into consideration the number of gallons each driver purchased.

The type of mean that considers an additional factor is called the *weighted mean*, and it is used when the values are not all equally represented.

Find the **weighted mean** of a variable X by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\bar{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum wX}{\sum w}$$

where $w_1, w_2, \ldots, w_n$ are the weights and $X_1, X_2, \ldots, X_n$ are the values.

Example 3–14 shows how the weighted mean is used to compute a grade point average. Since courses vary in their credit value, the number of credits must be used as weights.

EXAMPLE 3–14 Grade Point Average

A student received an A in English Composition I (3 credits), a C in Introduction to Psychology (3 credits), a B in Biology I (4 credits), and a D in Physical Education (2 credits). Assuming A = 4 grade points, B = 3 grade points, C = 2 grade points, D = 1 grade point, and F = 0 grade points, find the student's grade point average.

SOLUTION

Course	Credits (w)	Grade (X)
English Composition I	3	A (4 points)
Introduction to Psychology	3	C (2 points)
Biology I	4	B (3 points)
Physical Education	2	D (1 point)

$$\bar{X} = \frac{\Sigma wX}{\Sigma w} = \frac{3 \cdot 4 + 3 \cdot 2 + 4 \cdot 3 + 2 \cdot 1}{3 + 3 + 4 + 2} = \frac{32}{12} \approx 2.7$$

The grade point average is 2.7.

Unusual Stat

Of people in the United States, 45% live within 15 minutes of their best friend.

TABLE 3–1 Summary of Measures of Central Tendency

Measure	Definition	Symbol(s)
Mean	Sum of values, divided by total number of values	$\mu, \bar{X}$
Median	Middle point in data set that has been ordered	MD
Mode	Most frequent data value	None
Midrange	Lowest value plus highest value, divided by 2	MR

Table 3–1 summarizes the measures of central tendency.

Researchers and statisticians must know which measure of central tendency is being used and when to use each measure of central tendency. The properties and uses of the four measures of central tendency are summarized next.

Properties and Uses of Central Tendency

The Mean

1. The mean is found by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics, such as the variance.
4. The mean for the data set is unique and not necessarily one of the data values.
5. The mean cannot be computed for the data in a frequency distribution that has an open-ended class.
6. The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.

The Median

1. The median is used to find the center or middle value of a data set.
2. The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
3. The median is used for an open-ended distribution.
4. The median is affected less than the mean by extremely high or extremely low values.

The Mode

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode can be used when the data are nominal or categorical, such as religious preference, gender, or political affiliation.

4. The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

The Midrange

1. The midrange is easy to compute.
2. The midrange gives the midpoint.
3. The midrange is affected by extremely high or low values in a data set.

Distribution Shapes

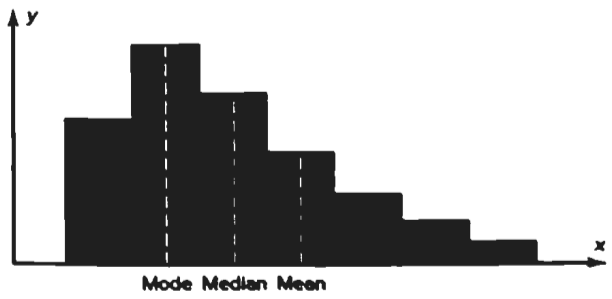
Frequency distributions can assume many shapes. The three most important shapes are positively skewed, symmetric, and negatively skewed. Figure 3–1 shows histograms of each.

In a **positively skewed** or **right-skewed distribution,** the majority of the data values fall to the left of the mean and cluster at the lower end of the distribution; the "tail" is to the right. Also, the mean is to the right of the median, and the mode is to the left of the median.

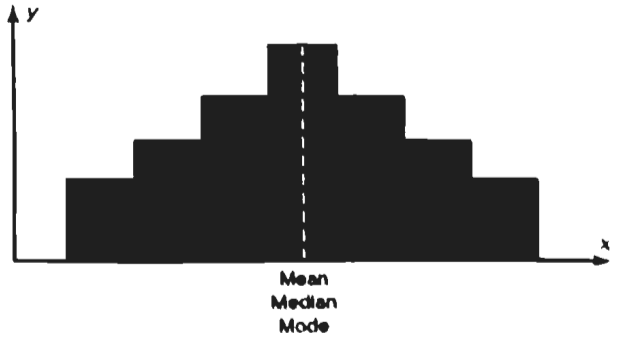
For example, if an instructor gave an examination and most of the students did poorly, their scores would tend to cluster on the left side of the distribution. A few high scores would constitute the tail of the distribution, which would be on the right side. Another example of a positively skewed distribution is the incomes of the population of the United States. Most of the incomes cluster about the low end of the distribution; those with high incomes are in the minority and are in the tail at the right of the distribution.

In a **symmetric distribution,** the data values are evenly distributed on both sides of the mean. In addition, when the distribution is unimodal, the mean, median, and mode are the same and are at the center of the distribution. Examples of symmetric distributions are IQ scores and heights of adult males.

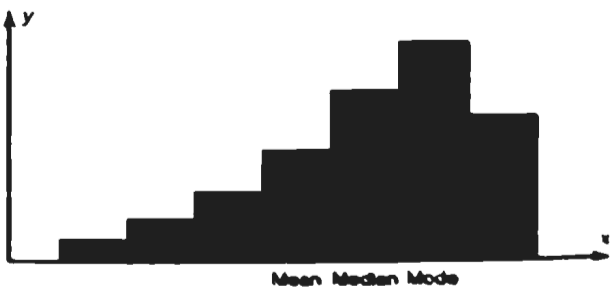
FIGURE 3–1
Types of Distributions



(a) Positively skewed or right-skewed



(b) Symmetric



(c) Negatively skewed or left-skewed

When the majority of the data values fall to the right of the mean and cluster at the upper end of the distribution, with the tail to the left, the distribution is said to be **negatively skewed** or **left-skewed.** Also, the mean is to the left of the median, and the mode is to the right of the median. As an example, a negatively skewed distribution results if the majority of students score very high on an instructor's examination. These scores will tend to cluster to the right of the distribution.

When a distribution is extremely skewed, the value of the mean will be pulled toward the tail, but the majority of the data values will be greater than the mean or less than the mean (depending on which way the data are skewed); hence, the median rather than the mean is a more appropriate measure of central tendency. An extremely skewed distribution can also affect other statistics.

A measure of skewness for a distribution is discussed in Exercise 48 in Section 3–2.

Applying the Concepts 3–1

Teacher Salaries

The following data from several years ago represent salaries (in dollars) from a school district in Greenwood, South Carolina.

10,000	11,000	11,000	12,500	14,300	17,500
18,000	16,600	19,200	21,560	16,400	107,000

1. First, assume you work for the school board in Greenwood and do not wish to raise taxes to increase salaries. Compute the mean, median, and mode, and decide which one would best support your position to not raise salaries.
2. Second, assume you work for the teachers' union and want a raise for the teachers. Use the best measure of central tendency to support your position.
3. Explain how outliers can be used to support one or the other position.
4. If the salaries represented every teacher in the school district, would the averages be parameters or statistics?
5. Which measure of central tendency can be misleading when a data set contains outliers?
6. When you are comparing the measures of central tendency, does the distribution display any skewness? Explain.

See page 184 for the answers.

Exercises 3–1

1. **Roller Coaster Speeds** The data show the heights in feet of 14 roller coasters. Find the mean, median, midrange, and mode for the data.

95	105	50	125	102	120	160
102	118	91	160	95	50	84

Source: RollerCoaster.com

2. **Airport Parking** The number of short-term parking spaces at 15 airports is shown. Find the mean, median, mode, and midrange for the data.

750	3400	1962	700	203
900	8662	260	1479	5905
9239	690	9822	1131	2516

Source: USA Today.

3. **Length of School Years** The lengths of school years in a sample of various countries in the world are shown. Find the mean, median, midrange, and mode of the data.

251	243	226	216	196	180

Source: U.S. News and World Report

4. **Observers in the Frogwatch Program** The number of observers in the Frogwatch USA program (a wildlife conservation program dedicated to helping conserve frogs and toads) for the top 10 states with the most observers is 484, 483, 422, 396, 378, 352, 338, 331, 318, and 302. The top 10 states with the most active watchers list these numbers of visits: 634, 464, 406, 267, 219, 194, 191, 150, 130, and 114. Find the mean, median, mode, and midrange for the data. Compare the measures of central tendency for these two groups of data.

Source: www.nwf.org/frogwatch

5. **Top Video Games** The following represent XBOX One Top Selling Games and units sold:

Titanfall	2,000,000
Call of Duty: Ghosts	1,790,000
Battlefield 4	1,340,000
Forza Motorsport 5	1,340,000
Tomb Rider: Definitive Edition	1,210,000
Dead Rising 3	1,100,000
Metal Gear Solid 5: Ground Zeroes	980,000
Assassins Creed 4: Black Flag	310,300
Madden NFL 25	303,000
Metro Redux	298,000

Find the mean, median, mode, and midrange.

Source: Statistic Brain Research Institute

6. **Earnings of Nonliving Celebrities** *Forbes* magazine prints an annual Top-Earning Nonliving Celebrities list (based on royalties and estate earnings). Find the mean, median, mode, and midrange for the data. Comment on the skewness. Figures represent millions of dollars.

Kurt Cobain	50	Ray Charles	10
Elvis Presley	42	Marilyn Monroe	8
Charles M. Schulz	35	Johnny Cash	8
John Lennon	24	J.R.R. Tolkien	7
Albert Einstein	20	George Harrison	7
Andy Warhol	19	Bob Marley	7
Theodore Geisel (Dr. Seuss)	10		

Source: articles.moneycentral.msn.com

7. **Paid Days Off** The data show the number of paid days off workers get in a sample of various countries of the world. Find the mean, median, midrange, and mode for the data.

38	29	10	34	28	30
30	26	19	20	25	

Source: Center for Economic and Policy Research

8. **Top-Paid CEOs** The data shown are the total compensation (in millions of dollars) for the 50 top-paid CEOs for a recent year. Compare the averages, and state which one you think is the best measure.

17.5	18.0	36.8	31.7	31.7
17.3	24.3	47.7	38.5	17.0
23.7	16.5	25.1	17.4	18.0
37.6	19.7	21.4	28.6	21.6
19.3	20.0	16.9	25.2	19.8
25.0	17.2	20.4	20.1	29.1
19.1	25.2	23.2	25.9	24.0
41.7	24.0	16.8	26.8	31.4
16.9	17.2	24.1	35.2	19.1
22.9	18.2	25.4	35.4	25.5

Source: USA TODAY

9. **Airline Passengers** The data show a sample of the number of passengers in millions that major airlines carried for a recent year. Find the mean, median, midrange, and mode for the data.

143.8	17.7	8.5	120.4	33.0	7.1	27.1
10.0	5.0	6.1	4.3	3.1	12.1	

Source: Airlines for America

10. **Foreign Workers** The number of foreign workers' certificates for the New England states and the northwestern states is shown. Find the mean, median, and mode for both areas and compare the results.

New England states	Northwestern states
6768	1870
3196	622
1112	620
819	23
1019	172
1795	112

Source: Department of Labor.

11. **Distances of Stars** Of the 25 brightest stars, the distances from earth (in light-years) for those with distances less than 100 light-years are found below. Find the mean, median, mode, and midrange for the data.

8.6	36.7	42.2	16.8	33.7	77.5	87.9
4.4	25.3	11.4	65.1	25.1	51.5	

Source: New York Times Almanac 2010.

12. **Contest Spelling Words** The last words given and spelled correctly at the National Spelling Bee for the past 21 years are spelled out below. Count the number of letters in each word, and find the mean, median, mode, and midrange for the data.

fibranne	cuonym	autochthonous
anupyretic	chiaroscurist	appoggiatura
lyceum	logorrhea	Umgrache
kamikaze	demarche	serrefine
antediluvian	succedaneum	guerdon
xanthosis	propaspicience	stromuhr
vivisepulture	pococurante	cymatrichous

Source: World Almanac 2012

13. Wind Speeds The data show the maximum wind speeds for a sample of 40 states. Find the mean and modal class for the data.

Class boundaries	Frequency
47.5–54.5	3
54.5–61.5	2
61.5–68.5	9
68.5–75.5	13
75.5–82.5	8
82.5–89.5	3
89.5–96.5	2

Source: NOAA

14. Hourly Compensation for Production Workers The hourly compensation costs (in U.S. dollars) for production workers in selected countries are represented below. Find the mean and modal class for the data.

Class	Frequency
2.48–7.48	7
7.49–12.49	3
12.50–17.50	1
17.51–22.51	7
22.52–27.52	5
27.53–32.53	5

Compare the mean of these grouped data to the U.S. mean of $21.97.

Source: New York Times Almanac.

15. Points in Rose Bowl Games The data show the number of points the winning team scored in the Rose Bowl. Find the mean and modal class for the data.

Class	Frequency
14–20	10
21–27	11
28–34	6
35–41	8
42–48	4
49–55	1

Source: The World Almanac and Book of Facts.

16. Percentage of Foreign-Born People The percentage of foreign-born population for each of the 50 states is represented here. Find the mean and modal class for the data. Do you think the mean is the best average for this set of data? Explain.

Percentage	Frequency
0.8–4.4	26
4.5–8.1	11
8.2–11.8	4
11.9–15.5	5
15.6–19.2	2
19.3–22.9	1
23.0–26.6	1

Source: World Almanac

17. Percentage of College-Educated Population over 25 Below are the percentages of the population over 25 years of age who have completed 4 years of college or more for the 50 states and the District of Columbia. Find the mean and modal class.

Percentage	Frequency
15.2–19.6	3
19.7–24.1	15
24.2–28.6	19
28.7–33.1	6
33.2–37.6	7
37.7–42.1	0
42.2–46.6	1

Source: New York Times Almanac.

18. Net Worth of Corporations These data represent the net worth (in millions of dollars) of 45 national corporations. Find the mean and modal class for the data.

Class limits	Frequency
10–20	2
21–31	8
32–42	15
43–53	7
54–64	10
65–75	3

19. Specialty Coffee Shops A random sample of 30 states shows the number of specialty coffee shops for a specific company. Find the mean and modal class for the data.

Class boundaries	Frequency
0.5–19.5	12
19.5–38.5	7
38.5–57.5	5
57.5–76.5	3
76.5–95.5	3

20. Commissions Earned This frequency distribution represents the commission earned (in dollars) by 100 salespeople employed at several branches of a large chain store. Find the mean and modal class for the data.

Class limits	Frequency
150–158	5
159–167	16
168–176	20
177–185	21
186–194	20
195–203	15
204–212	3

21. Children of U.S. Presidents The data show the number of children U.S. Presidents through Obama, had. Construct an ungrouped frequency distribution and find the mean and modal class.

0 5 6 0 3 4 1 4 10 8
7 0 6 1 0 3 4 5 4 8
7 3 5 2 1 2 1 5 3 3
0 0 2 2 6 1 2 3 2 2
4 4 2 6 1 2 2

Source: World Almanac and Book of Facts.

22. **Enrollments for Selected Independent Religiously Controlled 4-Year Colleges** Listed below are the enrollments for selected independent religiously controlled 4-year colleges that offer bachelor's degrees only. Construct a grouped frequency distribution with six classes and find the mean and modal class.

1013 1867 1268 1666 2309 1231 3005 2895 2166 1136
1532 1461 1750 1069 1723 1827 1155 1714 2391 2155
1412 1688 2471 1759 3008 2511 2577 1082 1067 1062
1319 1037 2400

Source: World Almanac.

23. **Automobile Selling Prices** Find the weighted mean price of three models of automobiles sold. The number and price of each model sold are shown in this list.

Model	Number	Price
A	8	$10,000
B	10	12,000
C	12	8,000

24. **Fat Grams** Using the weighted mean, find the average number of grams of fat per ounce of meat or fish that a person would consume over a 5-day period if he ate these:

Meat or fish	Fat (g/oz)
3 oz fried shrimp	3.33
3 oz veal cutlet (broiled)	3.00
2 oz roast beef (lean)	2.50
2.5 oz fried chicken drumstick	4.40
4 oz tuna (canned in oil)	1.75

Source: The World Almanac and Book of Facts.

25. **Diet Cola Preference** A recent survey of a new diet cola reported the following percentages of people who liked the taste. Find the weighted mean of the percentages.

Area	% Favored	Number surveyed
1	40	1000
2	30	3000
3	50	800

26. **Costs of Helicopters** The costs of three models of helicopters are shown here. Find the weighted mean of the costs of the models.

Model	Number sold	Cost
Sunscraper	9	$427,000
Skycoaster	6	365,000
High-flyer	12	725,000

27. **Final Grade** An instructor grades exams, 20%; term paper, 30%; final exam, 50%. A student had grades of 83, 72, and 90, respectively, for exams, term paper, and final exam. Find the student's final average. Use the weighted mean.

28. **Final Grade** Another instructor gives four 1-hour exams and one final exam, which counts as two 1-hour exams. Find a student's grade if she received 62, 83, 97, and 90 on the 1-hour exams and 82 on the final exam.

29. For these situations, state which measure of central tendency—mean, median, or mode—should be used.
    a. The most typical case is desired.
    b. The distribution is open-ended.
    c. There is an extreme value in the data set.
    d. The data are categorical.
    e. Further statistical computations will be needed.
    f. The values are to be divided into two approximately equal groups, one group containing the larger values and one containing the smaller values.

30. Describe which measure of central tendency—mean, median, or mode—was probably used in each situation.
    a. One-half of the factory workers make more than $5.37 per hour, and one-half make less than $5.37 per hour.
    b. The average number of children per family in the Plaza Heights Complex is 1.8.
    c. Most people prefer red convertibles over any other color.
    d. The average person cuts the lawn once a week.
    e. The most common fear today is fear of speaking in public.
    f. The average age of college professors is 42.3 years.

31. What types of symbols are used to represent sample statistics? Give an example. What types of symbols are used to represent population parameters? Give an example.

32. A local fast-food company claims that the average salary of its employees is $13.23 per hour. An employee states that most employees make minimum wage. If both are being truthful, how could both be correct?

Extending the Concepts

33. If the mean of five values is 64, find the sum of the values.

34. If the mean of five values is 8.2 and four of the values are 6, 10, 7, and 12, find the fifth value.

35. Find the mean of 10, 20, 30, 40, and 50.
    a. Add 10 to each value and find the mean.
    b. Subtract 10 from each value and find the mean.
    c. Multiply each value by 10 and find the mean.

d Divide each value by 10 and find the mean.

e Make a general statement about each situation.

36. Harmonic Mean The *harmonic mean* (HM) is defined as the number of values divided by the sum of the reciprocals of each value. The formula is

$$HM = \frac{n}{\Sigma(1/X)}$$

For example, the harmonic mean of 1, 4, 5, and 2 is

$$HM = \frac{4}{1/1 + 1/4 + 1/5 + 1/2} \approx 2.051$$

This mean is useful for finding the average speed. Suppose a person drove 100 miles at 40 miles per hour and returned driving 50 miles per hour. The average miles per hour is *not* 45 miles per hour, which is found by adding 40 and 50 and dividing by 2. The average is found as shown.

Since

$$\text{Time} = \text{distance} \div \text{rate}$$

then

$$\text{Time 1} = \frac{100}{40} = 2.5 \text{ hours to make the trip}$$

$$\text{Time 2} = \frac{100}{50} = 2 \text{ hours to return}$$

Hence, the total time is 4.5 hours, and the total miles driven are 200. Now, the average speed is

$$\text{Rate} = \frac{\text{distance}}{\text{time}} = \frac{200}{4.5} \approx 44.444 \text{ miles per hour}$$

This value can also be found by using the harmonic mean formula

$$HM = \frac{2}{1/40 + 1/50} \approx 44.444$$

Using the harmonic mean, find each of these.

a. A salesperson drives 300 miles round trip at 30 miles per hour going to Chicago and 45 miles per hour returning home. Find the average miles per hour.

b. A bus driver drives the 50 miles to West Chester at 40 miles per hour and returns driving 25 miles per hour. Find the average miles per hour.

c. A carpenter buys $500 worth of nails at $50 per pound and $500 worth of nails at $10 per pound. Find the average cost of 1 pound of nails.

37. Geometric Mean The *geometric mean* (GM) is defined as the *n*th root of the product of *n* values. The formula is

$$GM = \sqrt[n]{(X_1)(X_2)(X_3)\cdots(X_n)}$$

The geometric mean of 4 and 16 is

$$GM = \sqrt{(4)(16)} = \sqrt{64} = 8$$

The geometric mean of 1, 3, and 9 is

$$GM = \sqrt[3]{(1)(3)(9)} = \sqrt[3]{27} = 3$$

The geometric mean is useful in finding the average of percentages, ratios, indexes, or growth rates. For example, if a person receives a 20% raise after 1 year of service and a 10% raise after the second year of service, the average percentage raise per year is not 15 but 14.89%, as shown.

$$GM = \sqrt{(1.2)(1.1)} \approx 1.1489$$

or

$$GM = \sqrt{(120)(110)} \approx 114.89\%$$

His salary is 120% at the end of the first year and 110% at the end of the second year. This is equivalent to an average of 14.89%, since 114.89% − 100% = 14.89%.

This answer can also be shown by assuming that the person makes $10,000 to start and receives two raises of 20% and 10%.

$$\text{Raise 1} = 10{,}000 \cdot 20\% = \$2000$$
$$\text{Raise 2} = 12{,}000 \cdot 10\% = \$1200$$

His total salary raise is $3200. This total is equivalent to

$$\$10{,}000 \cdot 14.89\% = \$1489.00$$
$$\$11{,}489 \cdot 14.89\% = \$1710.71$$
$$\$3199.71 \approx \$3200$$

Find the geometric mean of each of these.

a. The growth rates of the Living Life Insurance Corporation for the past 3 years were 35, 24, and 18%.

b. A person received these percentage raises in salary over a 4-year period: 8, 6, 4, and 5%.

c. A stock increased each year for 5 years at these percentages: 10, 8, 12, 9, and 3%.

d. The price increases, in percentages, for the cost of food in a specific geographic region for the past 3 years were 1, 3, and 5.5%.

38. Quadratic Mean A useful mean in the physical sciences (such as voltage) is the *quadratic mean* (QM), which is found by taking the square root of the average of the squares of each value. The formula is

$$QM = \sqrt{\frac{\Sigma X^2}{n}}$$

The quadratic mean of 3, 5, 6, and 10 is

$$QM = \sqrt{\frac{3^2 + 5^2 + 6^2 + 10^2}{4}}$$
$$= \sqrt{42.5} \approx 6.519$$

Find the quadratic mean of 8, 6, 3, 5, and 4.

39. Median for Grouped Data An approximate median can be found for data that have been grouped into a frequency distribution. First it is necessary to find the median class. This is the class that contains the median value. That is the *n*/2 data value. Then it is assumed that

the data values are evenly distributed throughout the median class. The formula is

$$\text{MD} = \frac{n/2 - \text{cf}}{f}(w) + L_m$$

where n = sum of frequencies
cf = cumulative frequency of class immediately preceding the median class
w = width of median class
f = frequency of median class
L_m = lower boundary of median class

Using this formula, find the median for data in the frequency distribution of Exercise 16. ■

Technology Step by Step

EXCEL Step by Step

Finding Measures of Central Tendency

Example XL3–1

Find the mean, mode, and median of the data from Example 3–7. The data represent the population of licensed nuclear reactors in the United States for a recent 15-year period.

104	104	104	104	104
107	109	109	109	110
109	111	112	111	109

1. On an Excel worksheet enter the numbers in cells A2–A16. Enter a label for the variable in cell A1.

On the same worksheet as the data:

2. Compute the mean of the data: key in **=AVERAGE(A2:A16)** in a blank cell.
3. Compute the mode of the data: key in **=MODE(A2:A16)** in a blank cell.
4. Compute the median of the data: key in **=MEDIAN(A2:A16)** in a blank cell.

These and other statistical functions can also be accessed without typing them into the worksheet directly.

1. Select the Formulas tab from the toolbar and select the Insert Function Icon.
2. Select the Statistical category for statistical functions.
3. Scroll to find the appropriate function and click [OK].

(Excel reports only the first mode in a bimodal or multimodal distribution.) ■

3–2 Measures of Variation

In statistics, to describe the data set accurately, statisticians must know more than the measures of central tendency. Consider Example 3–15.

OBJECTIVE 2

Describe data, using measures of variation, such as the range, variance, and standard deviation.

EXAMPLE 3–15 Comparison of Outdoor Paint

A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the mean of each group.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

SOLUTION

The mean for brand A is

$$\mu = \frac{\Sigma X}{N} = \frac{210}{6} = 35 \text{ months}$$

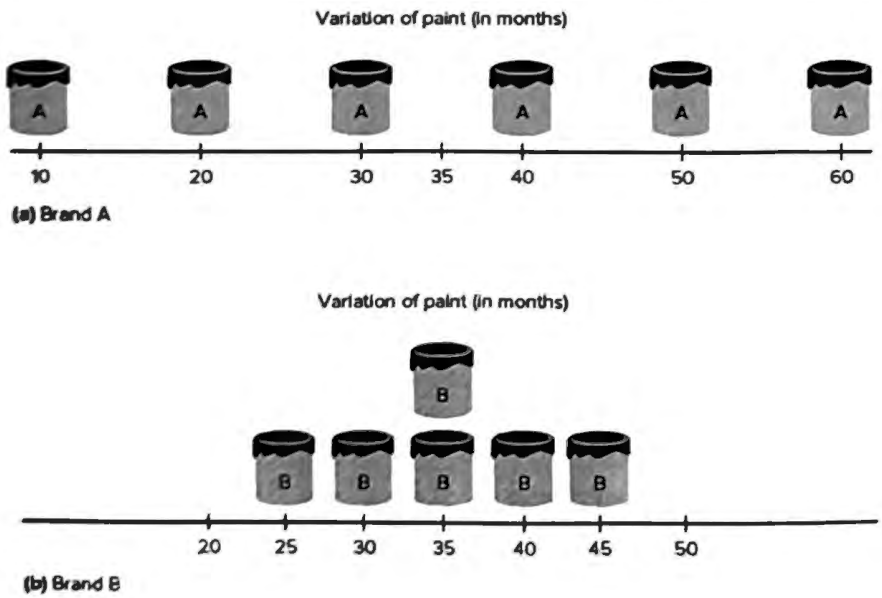
The mean for brand B is

$$\mu = \frac{\Sigma X}{N} = \frac{210}{6} = 35 \text{ months}$$

Since the means are equal in Example 3–15, you might conclude that both brands of paint last equally well. However, when the data sets are examined graphically, a somewhat different conclusion might be drawn. See Figure 3–2.

As Figure 3–2 shows, even though the means are the same for both brands, the spread, or variation, is quite different. Figure 3–2 shows that brand B performs more consistently; it is less variable. For the spread or variability of a data set, three measures are commonly used: *range, variance,* and *standard deviation.* Each measure will be discussed in this section.

FIGURE 3–2
Examining Data Sets Graphically



Range

The range is the simplest of the three measures and is defined now.

The range is the highest value minus the lowest value. The symbol R is used for the range.

$$R = \text{highest value} - \text{lowest value}$$

EXAMPLE 3–16 Comparison of Outdoor Paint

Find the ranges for the paints in Example 3–15.

SOLUTION

For brand A, the range is

$$R = 60 - 10 = 50 \text{ months}$$

For brand B, the range is

$$R = 45 - 25 = 20 \text{ months}$$

Make sure the range is given as a single number.

The range for brand A shows that 50 months separate the largest data value from the smallest data value. For brand B, 20 months separate the largest data value from the smallest data value, which is less than one-half of brand A's range.

One extremely high or one extremely low data value can affect the range markedly, as shown in Example 3–17.

EXAMPLE 3–17 Top-Grossing Movies

The data show a sample of the top-grossing movies in millions of dollars for a recent year. Find the range.

\$409 \$386 \$150 \$117 \$73 \$70

Source: The World Almanac and Book of Facts.

SOLUTION

Range $R = \$409 - \$70 = \$339$ million

The range for these data is quite large since it depends on the highest data value and the lowest data value. To have a more meaningful statistic to measure the variability, statisticians use measures called the *variance* and *standard deviation*.

Population Variance and Standard Deviation

Before these measures can be defined, it is necessary to know what *data variation* means. It is based on the difference or distance each data value is from the mean. This difference or distance is called a *deviation*. In the outdoor paint example, the mean for brand A paint is $\mu = 35$ months, and for a specific can, say, the can that lasted for 50 months, the deviation is $X - \mu$ or $50 - 35 = 15$. Hence, the deviation for that data value is 15 months. If you find the sum of the deviations for all data values about the mean (without rounding), this sum will always be zero. That is, $\Sigma(X - \mu) = 0$. (You can see this if you sum all the deviations for the paint example.)

To eliminate this problem, we sum the squares, that is, $\Sigma(X - \mu)^2$ and find the mean of these squares by dividing by N (the total number of data values), symbolically $\Sigma(X - \mu)^2/N$. This measure is called the *population variance* and is symbolized by σ^2, where σ is the symbol for Greek lowercase letter sigma.

Since this measure (σ^2) is in square units and the data are in regular units, statisticians take the square root of the variance and call it the *standard deviation.*

Formally defined,

The population variance is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 (σ is the Greek lowercase letter sigma).

The formula for the population variance is

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

where X = individual value
μ = population mean
N = population size

The population standard deviation is the square root of the variance. The symbol for the population standard deviation is σ.

The corresponding formula for the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

To find the variance and standard deviation for a data set, the following Procedure Table can be used.

Procedure Table

Finding the Population Variance and Population Standard Deviation

Step 1 Find the mean for the data.

$$\mu = \frac{\Sigma X}{N}$$

Step 2 Find the deviation for each data value.

$$X - \mu$$

Step 3 Square each of the deviations.

$$(X - \mu)^2$$

Step 4 Find the sum of the squares.

$$\Sigma(X - \mu)^2$$

Step 5 Divide by N to get the variance.

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

Step 6 Take the square root of the variance to get the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

Interesting Fact

The average American drives about 10,000 miles a year

Rounding Rule for the Standard Deviation The rounding rule for the standard deviation is the same as that for the mean. The final answer should be rounded to one more decimal place than that of the original data.

EXAMPLE 3–18 **Comparison of Outdoor Paint**

Find the variance and standard deviation for the data set for brand A paint in Example 3–15. The number of months brand A lasted before fading was

10, 60, 50, 30, 40, 20

SOLUTION

Step 1 Find the mean for the data.

$$\mu = \frac{\Sigma X}{N} = \frac{10 + 60 + 50 + 30 + 40 + 20}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each data value $(X - \mu)$.

$10 - 35 = -25$ $\quad 50 - 35 = 15$ $\quad 40 - 35 = 5$
$60 - 35 = 25$ $\quad 30 - 35 = -5$ $\quad 20 - 35 = -15$

Step 3 Square each result $(X - \mu)^2$.

$(-25)^2 = 625$ $\quad (15)^2 = 225$ $\quad (5)^2 = 25$
$(25)^2 = 625$ $\quad (-5)^2 = 25$ $\quad (-15)^2 = 225$

Step 4 Find the sum of the squares $\Sigma(X - \mu)^2$.

$$625 + 625 + 225 + 25 + 25 + 225 = 1750$$

Step 5 Divide the sum by N to get the variance $\frac{\Sigma(X - \mu)^2}{N}$.

$$\text{Variance} = 1750 \div 6 \approx 291.7$$

Step 6 Take the square root of the variance to get the standard deviation. Hence, the standard deviation equals $\sqrt{291.7}$, or 17.1. It is helpful to make a table.

A Values X	B $X - \mu$	C $(X - \mu)^2$
10	−25	625
60	25	625
50	15	225
30	−5	25
40	5	25
20	−15	225
		1750

Column A contains the raw data X. Column B contains the differences $X - \mu$ obtained in step 2. Column C contains the squares of the differences obtained in step 3.

The preceding computational procedure reveals several things. First, the square root of the variance gives the standard deviation; and vice versa, squaring the standard deviation gives the variance. Second, the variance is actually the average of the square of the distance that each value is from the mean. Therefore, if the values are near the mean, the variance will be small. In contrast, if the values are far from the mean, the variance will be large.

Historical Note

Karl Pearson in 1892 and 1893 introduced the statistical concepts of the range and standard deviation.

You might wonder why the squared distances are used instead of the actual distances. As previously stated, the reason is that the sum of the distances will always be zero. To verify this result for a specific case, add the values in column B of the table in Example 3–18. When each value is squared, the negative signs are eliminated.

Finally, why is it necessary to take the square root? Again, the reason is that since the distances were squared, the units of the resultant numbers are the squares of the units of

the original raw data. Finding the square root of the variance puts the standard deviation in the same units as the raw data.

When you are finding the square root, always use its positive value, since the variance and standard deviation of a data set can never be negative.

EXAMPLE 3–19 Comparison of Outdoor Paint

Find the variance and standard deviation for brand B paint data in Example 3–15. The months brand B lasted before fading were

35, 45, 30, 35, 40, 25

SOLUTION

Step 1 Find the mean.

$$\mu = \frac{\Sigma X}{N} = \frac{35 + 45 + 30 + 35 + 40 + 25}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each value, and place the result in column B of the table.

$35 - 35 = 0$ $\quad 45 - 35 = 10$ $\quad 30 - 35 = -5$
$35 - 35 = 0$ $\quad 40 - 35 = 5$ $\quad 25 - 35 = -10$

Step 3 Square each result and place the squares in column C of the table.

A X	B $X - \mu$	C $(X - \mu)^2$
35	0	0
45	10	100
30	−5	25
35	0	0
40	5	25
25	−10	100

Step 4 Find the sum of the squares in column C.

$$\Sigma(X - \mu)^2 = 0 + 100 + 25 + 0 + 25 + 100 = 250$$

Step 5 Divide the sum by N to get the variance.

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N} = \frac{250}{6} = 41.7$$

Step 6 Take the square root to get the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}} = \sqrt{41.7} \approx 6.5$$

Hence, the standard deviation is 6.5.

Since the standard deviation of brand A is 17.1 (see Example 3–18) and the standard deviation of brand B is 6.5, the data are more variable for brand A. *In summary, when the means are equal, the larger the variance or standard deviation is, the more variable the data are.*

Sample Variance and Standard Deviation

When computing the variance for a sample, one might expect the following expression to be used:

$$\frac{\Sigma(X - \bar{X})^2}{n}$$

where $\bar{X}$ is the sample mean and n is the sample size. *This formula is not usually used, however, since in most cases the purpose of calculating the statistic is to estimate the*

corresponding parameter. For example, the sample mean $\bar{X}$ is used to estimate the population mean μ. The expression

$$\frac{\Sigma(X - \bar{X})^2}{n}$$

does not give the best estimate of the population variance because when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance. Therefore, instead of dividing by n, find the variance of the sample by dividing by $n - 1$, giving a slightly larger value and an *unbiased* estimate of the population variance.

Formula for the Sample Variance

The formula for the sample variance (denoted by s^2) is

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1}$$

where X = individual value
$\bar{X}$ = sample mean
n = sample size

To find the standard deviation of a sample, you must take the square root of the sample variance, which was found by using the preceding formula.

Formula for the Sample Standard Deviation

The formula for the sample standard deviation, denoted by s, is

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$$

where X = individual value
$\bar{X}$ = sample mean
n = sample size

The procedure for finding the sample variance and the sample standard deviation is the same as the procedure for finding the population variance and the population standard deviation *except* the sum of the squares is divided by $n - 1$ (sample size minus 1) instead of N (population size). Refer to the previous Procedure Table if necessary. The next example shows these steps.

EXAMPLE 3–20 Teacher Strikes

The number of public school teacher strikes in Pennsylvania for a random sample of school years is shown. Find the sample variance and the sample standard deviation.

9 10 14 7 8 3

Source: Pennsylvania School Board Association.

SOLUTION

Step 1 Find the mean of the data values.

$$\bar{X} = \frac{\Sigma X}{n} = \frac{9 + 10 + 14 + 7 + 8 + 3}{6} = \frac{51}{6} = 8.5$$

Step 2 Find the deviation for each data value $(X - \bar{X})$.

$9 - 8.5 = 0.5$ $10 - 8.5 = 1.5$ $14 - 8.5 = 5.5$
$7 - 8.5 = -1.5$ $8 - 8.5 = -0.5$ $3 - 8.5 = -5.5$

Step 3 Square each of the deviations $(X - \bar{X})^2$.

$$(0.5)^2 = 0.25 \quad (1.5)^2 = 2.25 \quad (5.5)^2 = 30.25$$
$$(-1.5)^2 = 2.25 \quad (-0.5)^2 = 0.25 \quad (-5.5)^2 = 30.25$$

Step 4 Find the sum of the squares.

$$\Sigma(X - \bar{X})^2 = 0.25 + 2.25 + 30.25 + 2.25 + 0.25 + 30.25 = 65.5$$

Step 5 Divide by $n - 1$ to get the variance.

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} = \frac{65.5}{6 - 1} = \frac{65.5}{5} = 13.1$$

Step 6 Take the square root of the variance to get the standard deviation.

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{13.1} \approx 3.6 \text{ (rounded)}$$

Here the sample variance is 13.1, and the sample standard deviation is 3.6.

Shortcut formulas for computing the variance and standard deviation are presented next and will be used in the remainder of the chapter and in the exercises. These formulas are mathematically equivalent to the preceding formulas and do not involve using the mean. They save time when repeated subtracting and squaring occur in the original formulas. They are also more accurate when the mean has been rounded.

Shortcut or Computational Formulas for s^2 and s

The shortcut formulas for computing the variance and standard deviation for data obtained from samples are as follows.

Variance	Standard deviation
$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}$	$s = \sqrt{\frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}}$

Note that ΣX^2 is not the same as $(\Sigma X)^2$. The notation ΣX^2 means to square the values first, then sum; $(\Sigma X)^2$ means to sum the values first, then square the sum.

Example 3–21 explains how to use the shortcut formulas.

EXAMPLE 3–21 Teacher Strikes

The number of public school teacher strikes in Pennsylvania for a random sample of school years is shown. Find the sample variance and sample standard deviation.

9, 10, 14, 7, 8, 3

SOLUTION

Step 1 Find the sum of the values:

$$\Sigma X = 9 + 10 + 14 + 7 + 8 + 3 = 51$$

Step 2 Square each value and find the sum:

$$\Sigma X^2 = 9^2 + 10^2 + 14^2 + 7^2 + 8^2 + 3^2 = 499$$

Step 3 Substitute in the formula and solve:

$$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}$$
$$= \frac{6(499) - 51^2}{6(6-1)}$$

$$= \frac{2994 - 2601}{6(5)}$$

$$= \frac{393}{30}$$

$$= 13.1$$

The variance is 13.1.

$$s = \sqrt{13.1} \approx 3.6 \text{ (rounded)}$$

Hence, the sample variance is 13.1, and the sample standard deviation is 3.6. Notice that these are the same results as the results in Example 3–20.

Variance and Standard Deviation for Grouped Data

The procedure for finding the variance and standard deviation for grouped data is similar to that for finding the mean for grouped data, and it uses the midpoints of each class.

This procedure uses the shortcut formula, and X_m is the symbol for the class midpoint.

Shortcut or Computational Formula for s^2 and s for Grouped Data

Sample variance:

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)}$$

Sample standard deviation

$$s = \sqrt{\frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)}}$$

where X_m is the midpoint of each class and f is the frequency of each class.

The steps for finding the sample variance and sample standard deviation for grouped data are summarized in this Procedure Table.

Procedure Table

Finding the Sample Variance and Standard Deviation for Grouped Data

Step 1 Make a table as shown, and find the midpoint of each class.

A Class	B Frequency	C Midpoint	D $f \cdot X_m$	E $f \cdot X_m^2$

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

Step 4 Find the sums of columns B, D, and E. (The sum of column B is n. The sum of column D is $\Sigma f \cdot X_m$. The sum of column E is $\Sigma f \cdot X_m^2$.)

Step 5 Substitute in the formula and solve to get the variance.

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)}$$

Step 6 Take the square root to get the standard deviation.

EXAMPLE 3–22 Miles Run per Week

Find the sample variance and the sample standard deviation for the frequency distribution of the data shown. The data represent the number of miles that 20 runners ran during one week.

Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38

SOLUTION

Unusual Stat

At birth men outnumber women by 2%. By age 25, the number of men living is about equal to the number of women living. By age 65, there are 14% more women living than men.

Step 1 Make a table as shown, and find the midpoint of each class.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$	E $f \cdot X_m^2$
5.5–10.5	1	8		
10.5–15.5	2	13		
15.5–20.5	3	18		
20.5–25.5	5	23		
25.5–30.5	4	28		
30.5–35.5	3	33		
35.5–40.5	2	38		

Step 2 Multiply the frequency by the midpoint for each class, and place the products in column D.

$$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \qquad \ldots \qquad 2 \cdot 38 = 76$$

Step 3 Multiply the frequency by the square of the midpoint, and place the products in column E.

$$1 \cdot 8^2 = 64 \qquad 2 \cdot 13^2 = 338 \qquad \ldots \qquad 2 \cdot 38^2 = 2888$$

Step 4 Find the sums of columns B, D, and E. The sum of column B is n, the sum of column D is $\Sigma f \cdot X_m$, and the sum of column E is $\Sigma f \cdot X_m^2$. The completed table is shown.

A Class	B Frequency	C Midpoint	D $f \cdot X_m$	E $f \cdot X_m^2$
5.5–10.5	1	8	8	64
10.5–15.5	2	13	26	338
15.5–20.5	3	18	54	972
20.5–25.5	5	23	115	2,645
25.5–30.5	4	28	112	3,136
30.5–35.5	3	33	99	3,267
35.5–40.5	2	38	76	2,888
	$n = 20$		$\Sigma f \cdot X_m = 490$	$\Sigma f \cdot X_m^2 = 13{,}310$

Step 5 Substitute in the formula and solve for s^2 to get the variance.

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)}$$
$$= \frac{20(13{,}310) - 490^2}{20(20-1)}$$
$$= \frac{266{,}200 - 240{,}100}{20(19)}$$
$$= \frac{26{,}100}{380}$$
$$\approx 68.7$$

Step 6 Take the square root to get the standard deviation.

$$s \approx \sqrt{68.7} \approx 8.3$$

Be sure to use the number found in the sum of column B (i.e., the sum of the frequencies) for n. Do not use the number of classes.

The three measures of variation are summarized in Table 3–2.

Unusual Stat

The average number of times that a man cries in a month is 1.4.

TABLE 3–2 Summary of Measures of Variation

Measure	Definition	Symbol(s)
Range	Distance between highest value and lowest value	R
Variance	Average of the squares of the distance that each value is from the mean	σ^2, s^2
Standard deviation	Square root of the variance	σ, s

Uses of the Variance and Standard Deviation

1. As previously stated, variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.
2. The measures of variance and standard deviation are used to determine the consistency of a variable. For example, in the manufacture of fittings, such as nuts and bolts, the variation in the diameters must be small, or else the parts will not fit together.
3. The variance and standard deviation are used to determine the number of data values that fall within a specified interval in a distribution. For example, Chebyshev's theorem (explained later) shows that, for any distribution, at least 75% of the data values will fall within 2 standard deviations of the mean.
4. Finally, the variance and standard deviation are used quite often in inferential statistics. These uses will be shown in later chapters of this textbook.

Historical Note

Karl Pearson devised the coefficient of variation to compare the deviations of two different groups such as the heights of men and women.

Coefficient of Variation

Whenever two samples have the same units of measure, the variance and standard deviation for each can be compared directly. For example, suppose an automobile dealer wanted to compare the standard deviation of miles driven for the cars she received as trade-ins on new cars. She found that for a specific year, the standard deviation for Buicks was 422 miles and the standard deviation for Cadillacs was 350 miles. She could say that the variation in mileage was greater in the Buicks. But what if a manager wanted to compare the standard deviations of two different variables, such as the

number of sales per salesperson over a 3-month period and the commissions made by these salespeople?

A statistic that allows you to compare standard deviations when the units are different, as in this example, is called the *coefficient of variation*.

The coefficient of variation, denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples,	For populations,
$\text{CVar} = \frac{s}{\bar{X}} \cdot 100$	$\text{CVar} = \frac{\sigma}{\mu} \cdot 100$

EXAMPLE 3–23 Sales of Automobiles

The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

SOLUTION

The coefficients of variation are

$$\text{CVar} = \frac{s}{\bar{X}} \cdot 100 = \frac{5}{87} \cdot 100 \approx 5.7\% \quad \text{sales}$$

$$\text{CVar} = \frac{773}{5225} \cdot 100 \approx 14.8\% \quad \text{commissions}$$

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

EXAMPLE 3–24 Roller Coasters

The mean speed for the five fastest wooden roller coasters is 69.16 miles per hour, and the variance is 2.76. The mean height for the five tallest roller coasters is 177.80 feet, and the variance is 157.70. Compare the variations of the two data sets.

Source: Ultimate RollerCoaster.com

SOLUTION

$$\text{CVar} = \frac{\sqrt{2.76}}{69.16} \cdot 100 \approx 2.4\%$$

$$\text{CVar} = \frac{\sqrt{157.70}}{177.80} \cdot 100 \approx 7.1\%$$

The variation in the speeds is slightly larger than the variation in the heights of the roller coasters.

Range Rule of Thumb

The range can be used to approximate the standard deviation. The approximation is called the **range rule of thumb.**

The Range Rule of Thumb

A rough estimate of the standard deviation is

$$s \approx \frac{\text{range}}{4}$$

In other words, if the range is divided by 4, an approximate value for the standard deviation is obtained. For example, the standard deviation for the data set 5, 8, 8, 9, 10, 12, and 13 is 2.7, and the range is 13 − 5 = 8. The range rule of thumb is $s \approx 2$. The range rule of thumb in this case underestimates the standard deviation somewhat; however, it is in the ballpark.

A note of caution should be mentioned here. The range rule of thumb is only an *approximation* and should be used when the distribution of data values is unimodal and roughly symmetric.

The range rule of thumb can be used to estimate the largest and smallest data values of a data set. The smallest data value will be approximately 2 standard deviations below the mean, and the largest data value will be approximately 2 standard deviations above the mean of the data set. The mean for the previous data set is 9.3; hence,

$$\text{Smallest data value} = \overline{X} - 2s = 9.3 - 2(2.7) = 3.9$$

$$\text{Largest data value} = \overline{X} + 2s = 9.3 + 2(2.7) = 14.7$$

Notice that the smallest data value was 5, and the largest data value was 13. Again, these are rough approximations. For many data sets, almost all data values will fall within 2 standard deviations of the mean. Better approximations can be obtained by using Chebyshev's theorem and the empirical rule. These are explained next.

Chebyshev's Theorem

As stated previously, the variance and standard deviation of a variable can be used to determine the spread, or dispersion, of a variable. That is, the larger the variance or standard deviation, the more the data values are dispersed. For example, if two variables measured in the same units have the same mean, say, 70, and the first variable has a standard deviation of 1.5 while the second variable has a standard deviation of 10, then the data for the second variable will be more spread out than the data for the first variable. *Chebyshev's theorem*, developed by the Russian mathematician Chebyshev (1821–1894), specifies the proportions of the spread in terms of the standard deviation.

Chebyshev's theorem The proportion of values from a data set that will fall within k standard deviations of the mean will be at least $1 - 1/k^2$, where k is a number greater than 1 (k is not necessarily an integer).

This theorem states that at least three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean of the data set. This result is found by substituting $k = 2$ in the expression

$$1 - \frac{1}{k^2} \quad \text{or} \quad 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 75\%$$

For the example in which variable 1 has a mean of 70 and a standard deviation of 1.5, at least three-fourths, or 75%, of the data values fall between 67 and 73. These values are found by adding 2 standard deviations to the mean and subtracting 2 standard deviations from the mean, as shown:

$$70 + 2(1.5) = 70 + 3 = 73$$

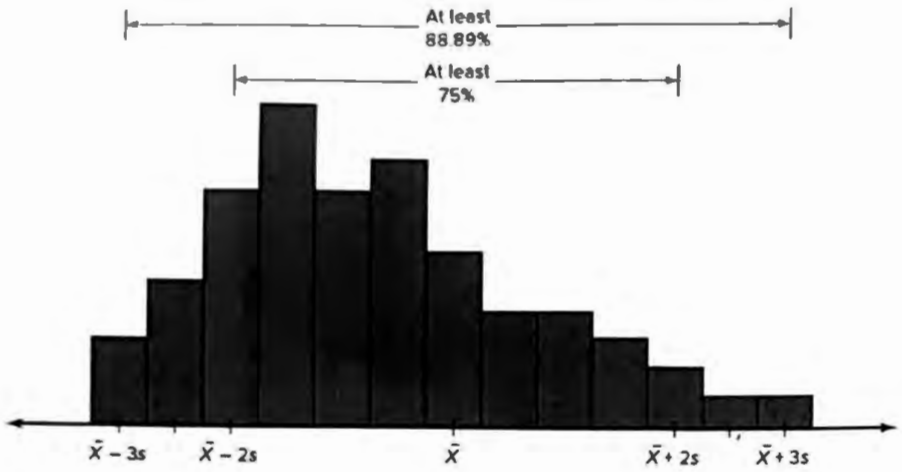
and

$$70 - 2(1.5) = 70 - 3 = 67$$

For variable 2, at least three-fourths, or 75%, of the data values fall between 50 and 90. Again, these values are found by adding and subtracting, respectively, 2 standard deviations to and from the mean.

$$70 + 2(10) = 70 + 20 = 90$$

FIGURE 3–3
Chebyshev's Theorem



and

$$70 - 2(10) = 70 - 20 = 50$$

Furthermore, the theorem states that at least eight-ninths, or 88.89%, of the data values will fall within 3 standard deviations of the mean. This result is found by letting $k = 3$ and substituting in the expression.

$$1 - \frac{1}{k^2} \quad \text{or} \quad 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} = 88.89\%$$

For variable 1, at least eight-ninths, or 88.89%, of the data values fall between 65.5 and 74.5, since

$$70 + 3(1.5) = 70 + 4.5 = 74.5$$

and

$$70 - 3(1.5) = 70 - 4.5 = 65.5$$

For variable 2, at least eight-ninths, or 88.89%, of the data values fall between 40 and 100.

In summary, then, Chebyshev's theorem states

- At least three-fourths, or 75%, of all data values fall within 2 standard deviations of the mean.
- At least eight-ninths, or 89%, of all data values fall within 3 standard deviations of the mean.

This theorem can be applied to any distribution regardless of its shape (see Figure 3–3). Examples 3–25 and 3–26 illustrate the application of Chebyshev's theorem.

EXAMPLE 3–25 Prices of Homes

The mean price of houses in a certain neighborhood is \$50,000, and the standard deviation is \$10,000. Find the price range for which at least 75% of the houses will sell.

SOLUTION

Chebyshev's theorem states that three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean. Thus,

$$\$50{,}000 + 2(\$10{,}000) = \$50{,}000 + \$20{,}000 = \$70{,}000$$

and

$$\$50{,}000 - 2(\$10{,}000) = \$50{,}000 - \$20{,}000 = \$30{,}000$$

Hence, at least 75% of all homes sold in the area will have a price range from \$30,000 to \$70,000.

Chebyshev's theorem can be used to find the minimum percentage of data values that will fall between any two given values. The procedure is shown in Example 3–26

EXAMPLE 3–26 **Travel Allowances**

A survey of local companies found that the mean amount of travel allowance for couriers was \$0.25 per mile. The standard deviation was \$0.02. Using Chebyshev's theorem, find the minimum percentage of the data values that will fall between \$0.20 and \$0.30.

SOLUTION

Step 1 Subtract the mean from the larger value.

$$\$0.30 - \$0.25 = \$0.05$$

Step 2 Divide the difference by the standard deviation to get k.

$$k = \frac{0.05}{0.02} = 2.5$$

Step 3 Use Chebyshev's theorem to find the percentage.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2.5^2} = 1 - \frac{1}{6.25} = 1 - 0.16 = 0.84 \quad \text{or} \quad 84\%$$

Hence, at least 84% of the data values will fall between \$0.20 and \$0.30.

The Empirical (Normal) Rule

Chebyshev's theorem applies to any distribution regardless of its shape. However, when a distribution is *bell-shaped* (or what is called *normal*), the following statements, which make up the **empirical rule,** are true.

- Approximately 68% of the data values will fall within 1 standard deviation of the mean.
- Approximately 95% of the data values will fall within 2 standard deviations of the mean.
- Approximately 99.7% of the data values will fall within 3 standard deviations of the mean.

For example, suppose that the scores on a national achievement exam have a mean of 480 and a standard deviation of 90. If these scores are normally distributed, then approximately 68% will fall between 390 and 570 (480 + 90 = 570 and 480 − 90 = 390). Approximately 95% of the scores will fall between 300 and 660 (480 + 2 · 90 = 660 and 480 − 2 · 90 = 300). Approximately 99.7% will fall between 210 and 750 (480 + 3 · 90 = 750 and 480 − 3 · 90 = 210). See Figure 3–4. (The empirical rule is explained in greater detail in Chapter 6.)

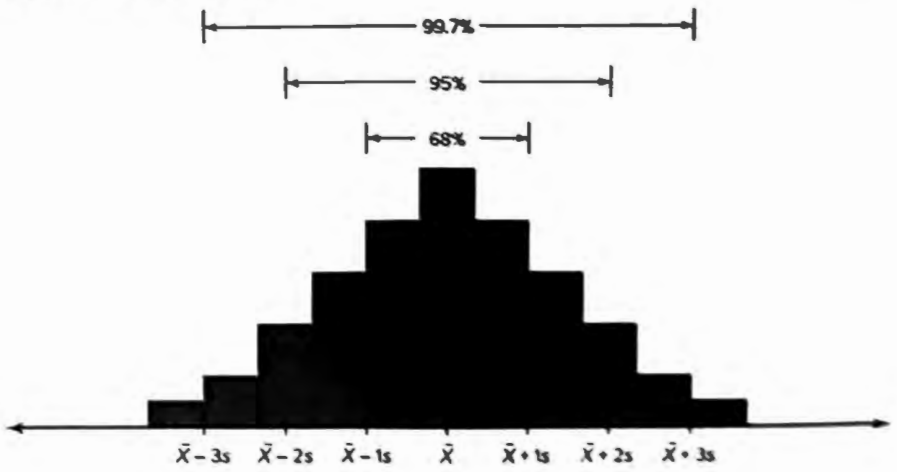


FIGURE 3–4
The Empirical Rule

Because the empirical rule requires that the distribution be approximately bell-shaped, the results are more accurate than those of Chebyshev's theorem, which applies to all distributions.

Linear Transformation of Data

In statistics, sometimes it is necessary to transform the data values into other data values. For example, if you are using temperature values collected from Canada, these values will be given using the Celsius temperature scale. If the study is to be used in the United States, you might want to change the data values to the Fahrenheit temperature scale. This change is called the *linear transformation of the data.* The question then arises, How does a linear transformation of the data values affect the mean and standard deviation of the data values?

Let's look at an example. Suppose you own a small store with five employees. Their hourly salaries are

$10 $13 $10 $11 $16

The mean of the salaries is $\overline{X} = \$12$, and the standard deviation is 2.550. Suppose you decide after a profitable year to give each employee a raise of $1.00 per hour. The new salaries would be

$11 $14 $11 $12 $17

The mean of the new salaries is $\overline{X} = \$13$, and the standard deviation of the new salaries is 2.550. Notice that the value of the mean increases by the amount that was added to each data value, and the standard deviation does not change.

Suppose that the five employees worked the number of hours per week shown here.

15 12 18 20 10

The mean for the number of hours for the original data is $\overline{X} = 15$, and the standard deviation for the number of hours is 4.123. You next decide to double the amount of each employee's hours for December. How does this affect the mean and standard deviation of the variable?

If each data value is multiplied by 2, the new data set is

30 24 36 40 20

The mean of the new data set is $\overline{X} = 30$, and the standard deviation is 8.246. The values of the mean and standard deviation double.

Hence, when each data value is multiplied by a constant, the mean of the new data set will be equal to the constant times the mean of the original set, and the standard deviation of the new data set will be equal to the absolute value (positive value) of the constant times the standard deviation of the original data set.

Applying the Concepts 3–2

Blood Pressure

The table lists means and standard deviations. The mean is the number before the plus/minus, and the standard deviation is the number after the plus/minus. The results are from a study attempting to find the average blood pressure of older adults. Use the results to answer the questions.

	Normotensive		Hypertensive	
	Men (n = 1200)	Women (n = 1400)	Men (n = 1100)	Women (n = 1300)
Age	55 ± 10	55 ± 10	60 ± 10	64 ± 10
Blood pressure (mm Hg)				
Systolic	123 ± 9	121 ± 11	153 ± 17	156 ± 20
Diastolic	78 ± 7	76 ± 7	91 ± 10	88 ± 10

1. Apply Chebyshev's theorem to the systolic blood pressure of normotensive men. At least how many of the men in the study fall within 1 standard deviation of the mean?
2. At least how many of those men in the study fall within 2 standard deviations of the mean?

Assume that blood pressure is normally distributed among older adults. Answer the following questions, using the empirical rule instead of Chebyshev's theorem.

3. What are the ranges for the diastolic blood pressure (normotensive and hypertensive) of older women?
4. Do the normotensive, male, systolic blood pressure ranges overlap with the hypertensive, male, systolic blood pressure ranges?

See page 184 for the answers.

Exercises 3–2

1. What is the relationship between the variance and the standard deviation?

2. Why might the range *not* be the best estimate of variability?

3. What are the symbols used to represent the population variance and standard deviation?

4. What are the symbols used to represent the sample variance and standard deviation?

5. Why is the unbiased estimator of variance used?

6. The three data sets have the same mean and range, but is the variation the same? Prove your answer by computing the standard deviation. Assume the data were obtained from samples.
   a. 5, 7, 9, 11, 13, 15, 17
   b. 5, 6, 7, 11, 15, 16, 17
   c. 5, 5, 5, 11, 17, 17, 17

7. **Traveler Spending** The data show the traveler spending in billions of dollars for a recent year for a sample of the states. Find the range, variance, and standard deviation for the data.

   20.1 33.5 21.7 58.4 23.2 110.8 30.9
   24.0 74.8 60.0

   Source: U.S. Travel Agency.

8. **Cigarette Taxes** The increases (in cents) in cigarette taxes for 17 states in a 6-month period are

   60, 20, 40, 40, 45, 12, 34, 51, 30, 70, 42, 31, 69, 32, 8, 18, 50

   Find the range, variance, and standard deviation for the data. Use the range rule of thumb to estimate the standard deviation. Compare the estimate to the actual standard deviation.

   Source: Federation of Tax Administrators.

9. **Prices of Silver and Tin** The following data show the price of silver and the price of tin over a recent 9-year period. Find the range, variance, and standard deviation. Which data set is more variable?

Silver	Tin
23.80	13.40
31.21	12.83
35.24	15.75
20.20	12.40
14.69	8.37
15.00	11.29
13.41	8.99
11.57	5.65
7.34	4.83

   Source: Department of the Interior.

10. **Size of U.S. States** The total surface area (in square miles) for each of six selected eastern states is listed here.

28,995	PA	37,534	FL
31,361	NY	27,087	VA
20,966	ME	37,741	GA

    The total surface area for each of six selected western states is listed (in square miles).

72,964	AZ	70,763	NV
101,510	CA	62,161	OR
66,625	CO	54,339	UT

    Find the standard deviation for each data set. Which set is more variable?

    Source: New York Times Almanac.

11. **Multiple Births** The numbers of various multiple births in the United States for the past 10 years are listed. Find the range, variance, and standard deviation of the data sets. Which set of data is the most variable?

Triplets			Quadruplets			Quintuplets		
5877	7110	5937	345	468	369	46	85	91
6898	6118	6885	434	355	501	69	67	85
6208	6742	6750	418	506	439	68	77	86
	6742			512			67	

    Source: World Almanac 2012

12. **Starting Teachers' Salaries** Starting teachers' salaries (in equivalent U.S. dollars) for upper secondary education in selected countries are listed. Find the range, variance, and standard deviation for the data. Which set of data is

more variable? (The U.S. average starting salary at this time was $29,641.)

Europe		Asia	
Sweden	$48,704	Korea	$26,852
Germany	41,441	Japan	23,493
Spain	32,679	India	18,247
Finland	32,136	Malaysia	13,647
Denmark	30,384	Philippines	9,857
Netherlands	29,326	Thailand	5,862
Scotland	27,789		

Source: World Almanac

13. **Ages of U.S. Astronaut Candidates** The average age of U.S. astronaut candidates in the past has been 34, but candidates have ranged in age from 26 to 46. Use the range rule of thumb to estimate the standard deviation of the applicants' ages.

Source: www.nasa.gov

14. **Times Spent in Rush-Hour Traffic** A sample of 12 drivers shows the time that they spent (in minutes) stopped in rush-hour traffic on a specific snowy day last winter. Find the range, variance, and standard deviation for the data.

52	56	53
61	49	51
53	58	53
60	71	58

15. **Laws Passed** The data show the number of public laws passed by the U.S. Congress for a sample of recent years. Find the range, variance, and standard deviation for the data.

283 394 383 580 498 460 377 482

Source: Congressional Record

16. **Passenger Vehicle Deaths** The number of people killed in each state from passenger vehicle crashes for a specific year is shown. Find the range, variance, and standard deviation for the data.

778	309	1110	324	705
1067	826	76	205	152
218	492	65	186	712
193	262	452	875	82
730	1185	2707	1279	390
305	123	948	343	602
69	451	951	104	985
155	450	2080	565	875
414	981	2786	82	793
214	130	396	620	797

Source: National Highway Traffic Safety Administration

17. **Annual Precipitation Days** The number of annual precipitation days for one-half of the 50 largest U.S. cities is listed below. Find the range, variance, and standard deviation of the data.

135	128	136	78	116	77	111	79	44	97
116	123	88	107	26	82	156	133	107	35
112	98	45	122	125					

18. Use the data from Exercises 7, 15, and 17 (unemployment, pressures, precipitation days) and compare the standard deviation with that obtained by the range rule of thumb (R/4.) Comment on the results.

19. **Pupils Per Teacher** The following frequency distribution shows the average number of pupils per teacher in the 50 states of the United States. Find the variance and standard deviation for the data.

Class limits	Frequency
9–11	2
12–14	20
15–17	18
18–20	7
21–23	2
24–26	1
	50

Source: U.S. Department of Education.

20. **Automotive Fuel Efficiency** Thirty automobiles were tested for fuel efficiency (in miles per gallon). This frequency distribution was obtained. Find the variance and standard deviation for the data.

Class boundaries	Frequency
7.5–12.5	3
12.5–17.5	5
17.5–22.5	15
22.5–27.5	5
27.5–32.5	2

21. **Murders in Cities** The data show the number of murders in 25 selected cities. Find the variance and standard deviation for the data.

Class limits	Frequency
34–96	13
97–159	2
160–222	0
223–285	5
286–348	1
349–411	1
412–474	0
475–537	1
538–600	2

22. **Reaction Times** In a study of reaction times to a specific stimulus, a psychologist recorded these data (in seconds). Find the variance and standard deviation for the data.

Class limits	Frequency
2.1–2.7	12
2.8–3.4	13
3.5–4.1	7
4.2–4.8	5
4.9–5.5	2
5.6–6.2	1

23. FM Radio Stations A random sample of 30 states shows the number of low-power FM radio stations for each state. Find the variance and standard deviation for the data.

Class limits	Frequency
1–9	5
10–18	7
19–27	10
28–36	3
37–45	3
46–54	2

Source: Federal Communications Commission.

24. Murder Rates The data represent the murder rate per 100,000 individuals in a sample of selected cities in the United States. Find the variance and standard deviation for the data.

Class limits	Frequency
5–11	8
12–18	5
19–25	7
26–32	1
33–39	1
40–46	3

Source: FBI and U.S. Census Bureau.

25. Waterfall Heights The frequency distribution shows a sample of the waterfall heights, in feet, of 28 waterfalls. Find the variance and standard deviation for the data.

Class boundaries	Frequency
52.5–185.5	8
185.5–318.5	11
318.5–451.5	2
451.5–584.5	1
584.5–717.5	4
717.5–850.5	2

Source: National Geographic Society.

26. Baseball Team Batting Averages Team batting averages for major league baseball in 2015 are represented below. Find the variance and standard deviation for each league. Compare the results.

NL		AL	
0.242–0.246	3	0.244–0.249	3
0.247–0.251	6	0.250–0.255	6
0.252–0.256	1	0.256–0.261	2
0.257–0.261	11	0.262–0.267	1
0.262–0.266	11	0.268–0.273	3
0.267–0.271	1	0.274–0.279	0

Source: World Almanac.

27. Missing Work The average number of days that construction workers miss per year is 11. The standard deviation is 2.3. The average number of days that factory workers miss per year is 8 with a standard deviation of 1.8. Which class is more variable in terms of days missed?

28. Suspension Bridges The lengths (in feet) of the main span of the longest suspension bridges in the United States and the rest of the world are shown below. Which set of data is more variable?

United States 4205, 4200, 3800, 3500, 3478, 2800, 2800, 2310
World 6570, 5538, 5328, 4888, 4626, 4544, 4518, 3970

Source: World Almanac

29. Hospital Emergency Waiting Times The mean of the waiting times in an emergency room is 80.2 minutes with a standard deviation of 10.5 minutes for people who are admitted for additional treatment. The mean waiting time for patients who are discharged after receiving treatment is 120.6 minutes with a standard deviation of 18.3 minutes. Which times are more variable?

30. Ages of Accountants The average age of the accountants at Three Rivers Corp. is 26 years, with a standard deviation of 6 years; the average salary of the accountants is $31,000, with a standard deviation of $4000. Compare the variations of age and income.

31. Using Chebyshev's theorem, solve these problems for a distribution with a mean of 80 and a standard deviation of 10.

a. At least what percentage of values will fall between 60 and 100?

b. At least what percentage of values will fall between 65 and 95?

32. The mean of a distribution is 20 and the standard deviation is 2. Use Chebyshev's theorem.

a. At least what percentage of the values will fall between 10 and 30?

b. At least what percentage of the values will fall between 12 and 28?

33. In a distribution of 160 values with a mean of 72, at least 120 fall within the interval 67–77. Approximately what percentage of values should fall in the interval 62–82? Use Chebyshev's theorem.

34. Calories in Bagels The average number of calories in a regular-size bagel is 240. If the standard deviation is 38 calories, find the range in which at least 75% of the data will lie. Use Chebyshev's theorem.

35. Time Spent Online Americans spend an average of 3 hours per day online. If the standard deviation is 32 minutes, find the range in which at least 88.89% of the data will lie. Use Chebyshev's theorem.

Source: www.[illegible].com

36. **Solid Waste Production** The average college student produces 640 pounds of solid waste each year. If the standard deviation is approximately 85 pounds, within what weight limits will at least 88.89% of all students' garbage lie?

Source: Environmental Sustainability Committee, www.cs.mtu.edu

37. **Sale Price of Homes** The average sale price of one-family houses in the United States for January 2016 was $258,100. Find the range of values in which at least 75% of the sale prices will lie if the standard deviation is $48,500.

Source: YCharts.com

38. **Trials to Learn a Maze** The average of the number of trials it took a sample of mice to learn to traverse a maze was 12. The standard deviation was 3. Using Chebyshev's theorem, find the minimum percentage of data values that will fall in the range of 4–20 trials.

39. **Farm Sizes** The average farm in the United States in 2014 contained 504 acres. The standard deviation is 55.7 acres. Use Chebyshev's theorem to find the minimum percentage of data values that will fall in the range of 364.75 and 643.25 acres.

Source: World Almanac

40. **Citrus Fruit Consumption** The average U.S. yearly per capita consumption of citrus fruit is 26.8 pounds. Suppose that the distribution of fruit amounts consumed is bell-shaped with a standard deviation equal to 4.2 pounds. What percentage of Americans would you expect to consume more than 31 pounds of citrus fruit per year?

Source: USDA/Economic Research Service

41. **SAT Scores** The national average for mathematics SATs in 2014 was 538. Suppose that the distribution of scores was approximately bell-shaped and that the standard deviation was approximately 48. Within what boundaries would you expect 68% of the scores to fall? What percentage of scores would be above 634?

42. **Work Hours for College Faculty** The average full-time faculty member in a postsecondary degree-granting institution works an average of 53 hours per week.
   a. If we assume the standard deviation is 2.8 hours, what percentage of faculty members work more than 58.6 hours a week?
   b. If we assume a bell-shaped distribution, what percentage of faculty members work more than 58.6 hours a week?

Source: National Center for Education Statistics.

43. **Prices of Musical Instruments** The average price of an instrument at a small music store is $325. The standard deviation of the price is $52. If the owner decides to raise the price of all the instruments by $20, what will be the new mean and standard deviation of the prices?

44. **Hours of Employment** The mean and standard deviation of the number of hours the employees work in the music store per week are, respectively, 18.6 and 3.2 hours. If the owner increases the number of hours each employee works per week by 4 hours, what will be the new mean and standard deviation of the number of hours worked by the employees?

45. **Price of Pet Fish** The mean price of the fish in a pet shop is $2.17, and the standard deviation of the price is $0.55. If the owner decides to triple the prices, what will be the mean and standard deviation of the new prices?

46. **Bonuses** The mean and standard deviation of the bonuses that the employees of a company received 10 years ago were, respectively, $2,000 and $325. Today the amount of the bonuses is 5 times what it was 10 years ago. Find the mean and standard deviation of the new bonuses.

Extending the Concepts

47. **Serum Cholesterol Levels** For this data set, find the mean and standard deviation of the variable. The data represent the serum cholesterol levels of 30 individuals. Count the number of data values that fall within 2 standard deviations of the mean. Compare this with the number obtained from Chebyshev's theorem. Comment on the answer.

211	240	255	219	204
200	212	193	187	205
256	203	210	221	249
231	212	236	204	187
201	247	206	187	200
237	227	221	192	196

48. **Ages of Consumers** For this data set, find the mean and standard deviation of the variable. The data represent the ages of 30 customers who ordered a product advertised on television. Count the number of data values that fall within 2 standard deviations of the mean. Compare this with the number obtained from Chebyshev's theorem. Comment on the answer.

42	44	62	35	20
30	56	20	23	41
55	22	31	27	66
21	18	24	42	25
32	50	31	26	36
39	40	18	36	22

49. Using Chebyshev's theorem, complete the table to find the minimum percentage of data values that fall within k standard deviations of the mean.

k	1.5	2	2.5	3	3.5
Percent					

50. Use this data set: 10, 20, 30, 40, 50
    a. Find the standard deviation.
    b. Add 5 to each value, and then find the standard deviation.
    c. Subtract 5 from each value and find the standard deviation.
    d. Multiply each value by 5 and find the standard deviation.
    e. Divide each value by 5 and find the standard deviation.
    f. Generalize the results of parts *b* through *e*.
    g. Compare these results with those in Exercise 35 of Exercises 3–1.

51. **Mean Deviation** The mean deviation is found by using this formula:

$$\text{Mean deviation} = \frac{\Sigma|X - \bar{X}|}{n}$$

where X = value
$\bar{X}$ = mean
n = number of values
$||$ = absolute value

Find the mean deviation for these data.

5, 9, 10, 11, 11, 12, 15, 18, 20, 22

52. **Pearson Coefficient of Skewness** A measure to determine the skewness of a distribution is called the *Pearson coefficient (PC) of skewness*. The formula is

$$\text{PC} = \frac{3(\bar{X} - \text{MD})}{s}$$

The values of the coefficient usually range from −3 to +3. When the distribution is symmetric, the coefficient is zero; when the distribution is positively skewed, it is positive; and when the distribution is negatively skewed, it is negative.

Using the formula, find the coefficient of skewness for each distribution, and describe the shape of the distribution.

a. Mean = 10, median = 8, standard deviation = 3.
b. Mean = 42, median = 45, standard deviation = 4.
c. Mean = 18.6, median = 18.6, standard deviation = 1.5.
d. Mean = 98, median = 97.6, standard deviation = 4.

53. All values of a data set must be within $s\sqrt{n-1}$ of the mean. If a person collected 25 data values that had a mean of 50 and a standard deviation of 3 and you saw that one data value was 67, what would you conclude?

Technology Step by Step

EXCEL Step by Step

Finding Measures of Variation

Example XL3–2

Find the sample variance, sample standard deviation, and range of the data from Example 3–20.

9 10 14 7 8 3

1. On an Excel worksheet enter the data in cells A2–A7. Enter a label for the variable in cell A1.
2. In a blank cell enter = **VAR.S(A2:A7)** for the sample variance.
3. In a blank cell enter = **STDEV.S(A2:A7)** for the sample standard deviation.
4. For the range, compute the difference between the maximum and the minimum values by entering = **Max(A2:A7)-Min(A2:A7)**.

Note: The command for computing the population variance is VAR.P and for the population standard deviation is STDEV.P

These and other statistical functions can also be accessed without typing them into the worksheet directly.

1. Select the Formulas tab from the Toolbar and select the Insert Function Icon,
2. Select the Statistical category for statistical functions.
3. Scroll to find the appropriate function and click [OK].

	A	B	C	D
1	Strikes			
2	9		Variance	13.1
3	10		Standard Deviation	3.619392214
4	14		Range	11
5	7			
6	8			
7	3			

3–3 Measures of Position

OBJECTIVE 1

Identify the position of a data value in a data set using various measures of position such as percentiles, deciles, and quartiles.

In addition to measures of central tendency and measures of variation, there are measures of position or location. These measures include standard scores, percentiles, deciles, and quartiles. They are used to locate the relative position of a data value in the data set. For example, if a value is located at the 80th percentile, it means that 80% of the values fall below it in the distribution and 20% of the values fall above it. The *median* is the value that corresponds to the 50th percentile, since one-half of the values fall below it and one-half of the values fall above it. This section discusses these measures of position.

Standard Scores

There is an old saying, "You can't compare apples and oranges." But with the use of statistics, it can be done to some extent. Suppose that a student scored 90 on a music test and 45 on an English exam. Direct comparison of raw scores is impossible, since the exams might not be equivalent in terms of number of questions, value of each question, and so on. However, a comparison of a relative standard similar to both can be made. This comparison uses the mean and standard deviation and is called a *standard score* or *z* score. (We also use *z* scores in later chapters.)

A standard score or *z* score tells how many standard deviations a data value is above or below the mean for a specific distribution of values. If a standard score is zero, then the data value is the same as the mean.

A *z* score or standard score for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is *z*. The formula is

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

For samples, the formula is

$$z = \frac{X - \bar{X}}{s}$$

For populations, the formula is

$$z = \frac{X - \mu}{\sigma}$$

The *z* score represents the number of standard deviations that a data value falls above or below the mean.

For the purpose of this section, it will be assumed that when we find *z* scores, the data were obtained from samples.

Interesting Fact

The average number of faces that a person learns to recognize and remember during his or her lifetime is 10,000.

EXAMPLE 3–27 Test Scores

A student scored 85 on an English test while the mean score of all the students was 76 and the standard deviation was 4. She also scored 42 on a French test where the class mean was 36 and the standard deviation was 3. Compare the relative positions on the two tests.

SOLUTION

First find the *z* scores. For the English test

$$z = \frac{X - \bar{X}}{s} = \frac{85 - 76}{4} = 2.25$$

For the French test

$$z = \frac{X - \bar{X}}{s} = \frac{42 - 36}{3} = 2.00$$

Since the z score for the English test is higher than the z score for the French test, her relative position in the English class is higher than her relative position in the French class.

Note that if the z score is positive, the score is above the mean. If the z score is 0, the score is the same as the mean. And if the z score is negative, the score is below the mean.

EXAMPLE 3–28 Marriage Ages

In a recent study, the mean age at which men get married is said to be 26.4 years with a standard deviation of 2 years. The mean age at which women marry is 23.5 years with a standard deviation of 2.3 years. Find the relative positions for a man who marries at age 24 and a woman who marries at age 22.

SOLUTION

$$\text{Man } z = \frac{X - \bar{X}}{s} = \frac{24 - 26.4}{2} = -1.2$$

$$\text{Woman } z = \frac{X - \bar{X}}{s} = \frac{22 - 23.5}{2.3} \approx -0.65$$

In this case, the woman's age at marriage is relatively higher than the man's age at marriage.

When all data for a variable are transformed into z scores, the resulting distribution will have a mean of 0 and a standard deviation of 1. A z score, then, is actually the number of standard deviations each value is from the mean for a specific distribution. In Example 3-27, the English test score was 2.25 standard deviations above the mean, while the French test score was 2 standard deviations above the mean. This will be explained in greater detail in Chapter 6.

Percentiles

Percentiles are position measures used in educational and health-related fields to indicate the position of an individual in a group.

Percentiles divide the data set into 100 equal groups.

Percentiles are symbolized by

$$P_1, P_2, P_3, \ldots, P_{99}$$

and divide the distribution into 100 groups.

Interesting Facts

The highest recorded temperature on earth was 136°F in Libya in 1922. The lowest recorded temperature on earth was −129°F in Antarctica in 1983.

TABLE 3–3 Percentile Ranks and Scaled Scores on the Test of English as a Foreign Language*

Scaled score	Section 1: Listening comprehension	Section 2: Structure and written expression	Section 3: Vocabulary and reading comprehension	Total scaled score	Percentile rank
68	99	98			
66	98	96	98	660	99
64	96	94	96	640	97
62	92	90	93	620	94
60	87	84	88	600	89
−58	81	76	81	580	82
56	73	68	72	560	73
54	64	58	61	540	62
52	54	48	50	520	50
50	42	38	40	500	39
48	32	29	30	480	29
46	22	21	23	460	20
44	14	15	16	440	13
42	9	10	11	420	9
40	5	7	8	400	5
38	3	4	5	380	3
36	2	3	3	360	1
34	1	2	2	340	1
32		1	1	320	
30		1	1	300	
Mean	51.5	52.2	51.4	517	Mean
S.D.	7.1	7.9	7.5	68	S.D.

*Based on the total group of 1,178,193 examinees.

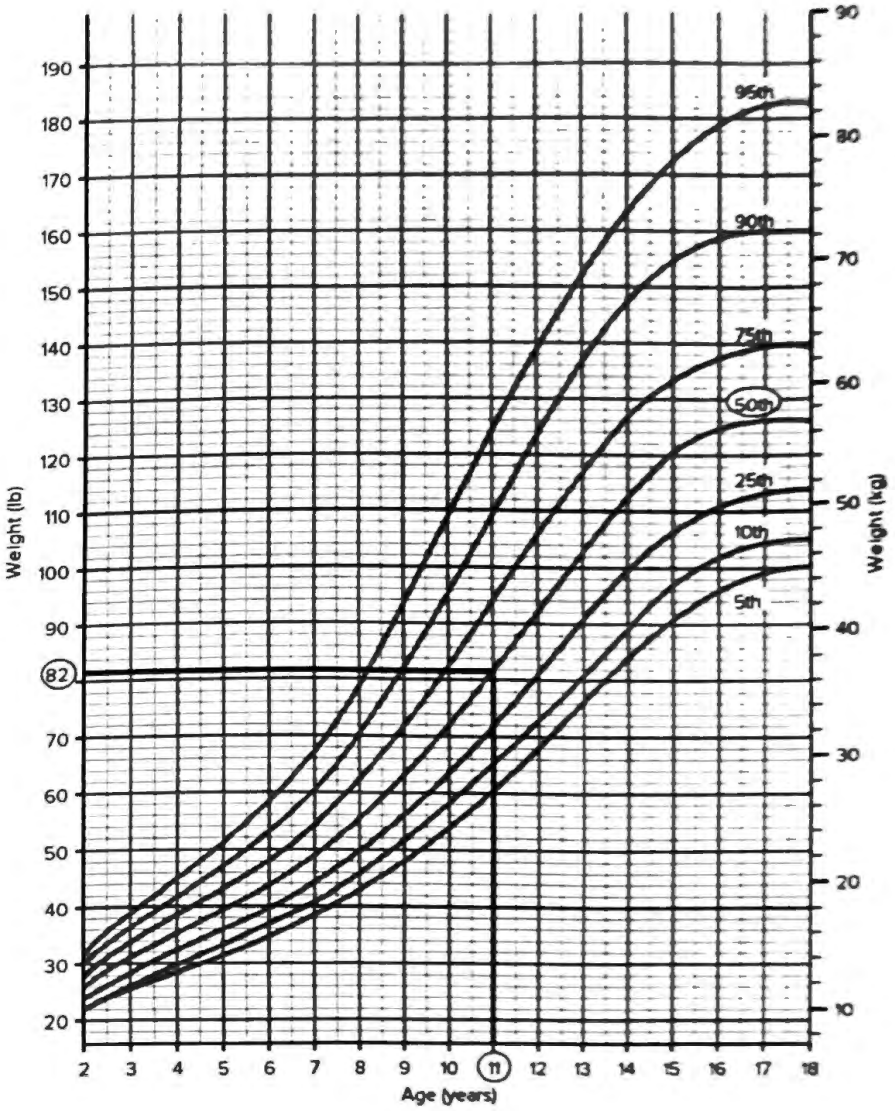
Source: Data from Educational Testing Service.

In many situations, the graphs and tables showing the percentiles for various measures such as test scores, heights, or weights have already been completed. Table 3–3 shows the percentile ranks for scaled scores on the Test of English as a Foreign Language. If a student had a scaled score of 58 for section 1 (listening and comprehension), that student would have a percentile rank of 81. Hence, that student did better than 81% of the students who took section 1 of the exam.

Figure 3–5 shows percentiles in graphical form of weights of girls from ages 2 to 18. To find the percentile rank of an 11-year-old who weighs 82 pounds, start at the 82-pound weight on the left axis and move horizontally to the right. Find 11 on the horizontal axis and move up vertically. The two lines meet at the 50th percentile curved line; hence, an 11-year-old girl who weighs 82 pounds is in the 50th percentile for her age group. If the lines do not meet exactly on one of the curved percentile lines, then the percentile rank must be approximated.

Percentiles are also used to compare an individual's test score with the national norm. For example, tests such as the National Educational Development Test (NEDT) are taken by students in ninth or tenth grade. A student's scores are compared with those of other

FIGURE 3–5
Weights of Girls by Age and Percentile Rankings
Source: Centers for Disease Control and Prevention



students locally and nationally by using percentile ranks. A similar test for elementary school students is called the California Achievement Test.

Percentiles are not the same as percentages. That is, if a student gets 72 correct answers out of a possible 100, she obtains a percentage score of 72. There is no indication of her position with respect to the rest of the class. She could have scored the highest, the lowest, or somewhere in between. On the other hand, if a raw score of 72 corresponds to the 64th percentile, then she did better than 64% of the students in her class.

Percentile graphs can be constructed as shown in Example 3–29 and Figure 3–6. Percentile graphs use the same values as the cumulative relative frequency graphs described in Section 2–2, except that the proportions have been converted to percents.

EXAMPLE 3–29 Systolic Blood Pressure

The frequency distribution for the systolic blood pressure readings (in millimeters of mercury, mm Hg) of 200 randomly selected college students is shown here. Construct a percentile graph.

A Class boundaries	B Frequency	C Cumulative frequency	D Cumulative percent
89.5–104.5	24		
104.5–119.5	62		
119.5–134.5	72		
134.5–149.5	26		
149.5–164.5	12		
164.5–179.5	4		
	200		

SOLUTION

Step 1 Find the cumulative frequencies and place them in column C.

Step 2 Find the cumulative percentages and place them in column D. To do this step, use the formula

$$\text{Cumulative } \% = \frac{\text{cumulative frequency}}{n} \cdot 100$$

For the first class,

$$\text{Cumulative } \% = \frac{24}{200} \cdot 100 = 12\%$$

The completed table is shown here.

A Class boundaries	B Frequency	C Cumulative frequency	D Cumulative percent
89.5–104.5	24	24	12
104.5–119.5	62	86	43
119.5–134.5	72	158	79
134.5–149.5	26	184	92
149.5–164.5	12	196	98
164.5–179.5	4	200	100
	200		

Step 3 Graph the data, using class boundaries for the x axis and the percentages for the y axis, as shown in Figure 3–6.

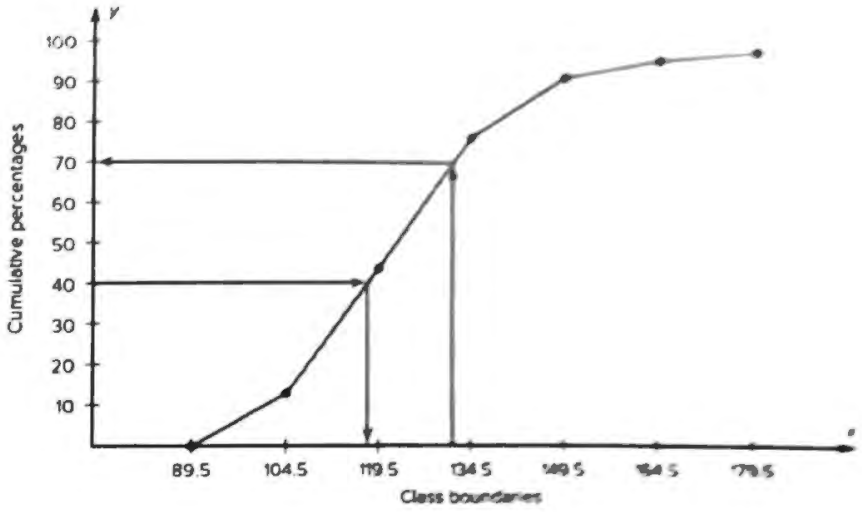
Once a percentile graph has been constructed, one can find the approximate corresponding percentile ranks for given blood pressure values and find approximate blood pressure values for given percentile ranks.

For example, to find the percentile rank of a blood pressure reading of 130, find 130 on the x axis of Figure 3–6 and draw a vertical line to the graph. Then move horizontally to the value on the y axis. Note that a blood pressure of 130 corresponds to approximately the 70th percentile.

If the value that corresponds to the 40th percentile is desired, start on the y axis at 40 and draw a horizontal line to the graph. Then draw a vertical line to the x axis and read the value. In Figure 3–6, the 40th percentile corresponds to a value of approximately 118. Thus, if a person has a blood pressure of 118, he or she is at the 40th percentile.

Finding values and the corresponding percentile ranks by using a graph yields only approximate answers. Several mathematical methods exist for computing percentiles for data. These methods can be used to find the approximate percentile rank of a data value

FIGURE 3–6
Percentile Graph for Example 3–29



or to find a data value corresponding to a given percentile. When the data set is large (100 or more), these methods yield better results. Examples 3–30 and 3–31 show these methods.

Percentile Formula

The percentile corresponding to a given value X is computed by using the following formula:

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100$$

EXAMPLE 3–30 Traffic Violations

The number of traffic violations recorded by a police department for a 10-day period is shown. Find the percentile rank of 16.

22 19 25 24 18 15 9 12 16 20

SOLUTION

Arrange the data in order from lowest to highest.

9 12 15 16 18 19 20 22 24 25

Then substitute into the formula.

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100$$

Since there are 3 numbers below the value of 16, the solution is

$$\text{Percentile} = \frac{3 + 0.5}{10} \cdot 100 = 35\text{th percentile}$$

Hence, the value of 16 is higher than 35% of the data values.

Note: One assumes that a value of 16, for instance, means theoretically any value between 15.5 and 16.5.

EXAMPLE 3–31 Traffic Violations

Using the data in Example 3–30, find the percentile rank of 24.

SOLUTION

There are 8 values below 24; thus,

$$\text{Percentile} = \frac{8 + 0.5}{10} \cdot 100 = 85\text{th percentile}$$

Therefore, the data value 24 is higher than 85% of the values in the data set.

The steps for finding a value corresponding to a given percentile are summarized in this Procedure Table.

Procedure Table

Finding a Data Value Corresponding to a Given Percentile

Step 1 Arrange the data in order from lowest to highest.

Step 2 Substitute into the formula

$$c = \frac{n \cdot p}{100}$$

where n = total number of values
p = percentile

Step 3A If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded-up value.

Step 3B If c is a whole number, use the value halfway between the cth and $(c + 1)$st values when counting up from the lowest value.

Examples 3–32 and 3–33 show a procedure for finding a value corresponding to a given percentile.

EXAMPLE 3–32 Traffic Violations

Using the data in Example 3–30, find the value corresponding to the 65th percentile.

SOLUTION

Step 1 Arrange the data in order from lowest to highest.

9 12 15 16 18 19 20 22 24 25

Step 2 Compute

$$c = \frac{n \cdot p}{100}$$

where n = total number of values
p = percentile

Thus,

$$c = \frac{10 \cdot 65}{100} = 6.5$$

Since c is not a whole number, round it up to the next whole number; in this case, it is $c = 7$.

Start at the lowest value and count over to the 7th value, which is 20. Hence, the value of 20 corresponds to the 65th percentile.

EXAMPLE 3–33 **Traffic Violations**

Using the data in Example 3–30, find the data value corresponding to the 30th percentile.

SOLUTION

Step 1 Arrange the data in order from lowest to highest.

9 12 15 16 18 19 20 22 24 25

Step 2 Substitute in the formula.

$$c = \frac{n \cdot p}{100} \qquad c = \frac{10 \cdot 30}{100} = 3$$

In this case, it is the 3rd and 4th values.

Step 3 Since c is a whole number, use the value halfway between the c and $c + 1$ values when counting up from the lowest. In this case, it is the third and fourth values.

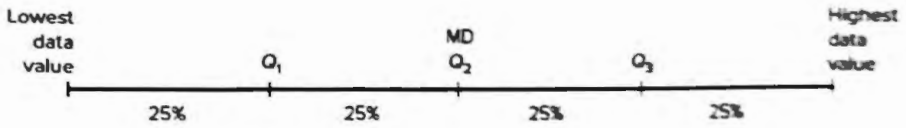
9 12 15 16 18 19 20 22 24 25
↑ ↑
3rd value 4th value

The halfway value is between 15 and 16. It is 15.5. Hence, 15.5 corresponds to the 30th percentile.

Quartiles and Deciles

Quartiles divide the distribution into four equal groups, denoted by Q_1, Q_2, Q_3.

Note that Q_1 is the same as the 25th percentile; Q_2 is the same as the 50th percentile, or the median; Q_3 corresponds to the 75th percentile, as shown:



Quartiles can be computed by using the formula given for computing percentiles on page 153. For Q_1 use $p = 25$. For Q_2 use $p = 50$. For Q_3 use $p = 75$. However, an easier method for finding quartiles is found in this Procedure Table.

Procedure Table

Finding Data Values Corresponding to Q_1, Q_2, and Q_3

Step 1 Arrange the data in order from lowest to highest.

Step 2 Find the median of the data values. This is the value for Q_2.

Step 3 Find the median of the data values that fall below Q_2. This is the value for Q_1.

Step 4 Find the median of the data values that fall above Q_2. This is the value for Q_3.

Example 3–34 shows how to find the values of Q_1, Q_2, and Q_3.

EXAMPLE 3–34 Traffic Violations

Using the data in Exercise 3–30, find Q_1, Q_2, and Q_3.

SOLUTION

Step 1 Arrange the data in order from lowest to highest.

9 12 15 16 18 19 20 22 24 25

Step 2 Find the median Q_2.

9 12 15 16 18 19 20 22 24 25

$$\text{MD} = \frac{18 + 19}{2} = 18.5$$

Step 3 Find the median of the data values below 18.5.

9 12 15 16 18

$Q_1 = 15$

Step 4 Find the median of the data values greater than 18.5.

19 20 22 24 25

$Q_3 = 22$

Hence, $Q_1 = 15$, $Q_2 = 18.5$, and $Q_3 = 22$.

In addition to dividing the data set into four groups, quartiles can be used as a rough measure of variability. This measure of variability which uses quartiles is called the *interquartile range* and is the range of the middle 50% of the data values.

The **interquartile range (IQR)** is the difference between the third and first quartiles.

$$\text{IQR} = Q_3 - Q_1$$

EXAMPLE 3–35 Traffic Violations

Find the interquartile range of the data set in Example 3–30.

SOLUTION

Find Q_1 and Q_3. This was done in Example 3–34. Now $Q_1 = 15$ and $Q_3 = 22$. Next subtract Q_1 from Q_3.

$$\text{IQR} = Q_3 - Q_1 = 22 - 15 = 7$$

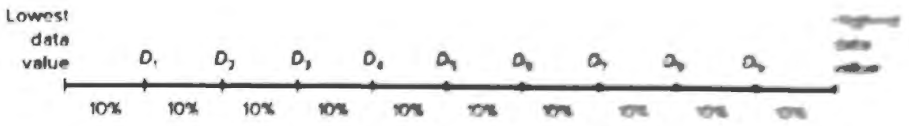
The interquartile range is equal to 7.

Like the standard deviation, the more variable the data set is, the larger the value of the interquartile range will be.

Deciles divide the distribution into 10 groups, as shown. They are denoted by D_1, D_2, etc.

Unusual Stat

Of the alcoholic beverages consumed in the United States, 85% is beer.



Note that D_1 corresponds to P_{10}; D_2 corresponds to P_{20}, etc. Deciles can be found by using the formulas given for percentiles. Taken altogether then, these are the relationships among percentiles, deciles, and quartiles.

Deciles are denoted by $D_1, D_2, D_3, \ldots, D_9$, and they correspond to $P_{10}, P_{20}, P_{30}, \ldots, P_{90}$.

Quartiles are denoted by Q_1, Q_2, Q_3 and they correspond to P_{25}, P_{50}, P_{75}.

The median is the same as P_{50} or Q_2 or D_5.

The position measures are summarized in Table 3–4

TABLE 3–4 Summary of Position Measures

Measure	Definition	Symbol(s)
Standard score or *z* score	Number of standard deviations that a data value is above or below the mean	z
Percentile	Position in hundredths that a data value holds in the distribution	P_n
Decile	Position in tenths that a data value holds in the distribution	D_n
Quartile	Position in fourths that a data value holds in the distribution	Q_n

Outliers

A data set should be checked for extremely high or extremely low values. These values are called *outliers*.

An outlier is an extremely high or an extremely low data value when compared with the rest of the data values.

An outlier can strongly affect the mean and standard deviation of a variable. For example, suppose a researcher mistakenly recorded an extremely high data value. This value would then make the mean and standard deviation of the variable much larger than they really were.

Since these measures (mean and standard deviation) are affected by outliers, they are called *nonresistant statistics*. The median and interquartile range are less affected by outliers, so they are called *resistant* statistics. Sometimes when a distribution is skewed or contains outliers, the median and interquartile range can be used to more accurately describe the data than the mean and standard deviation. Outliers can have an effect on other statistics as well.

There are several ways to check a data set for outliers. One method is shown in this Procedure Table.

Procedure Table

Procedure for Identifying Outliers

Step 1 Arrange the data in order from lowest to highest and find Q_1 and Q_3.

Step 2 Find the interquartile range: $IQR = Q_3 - Q_1$.

Step 3 Multiply the IQR by 1.5.

Step 4 Subtract the value obtained in step 3 from Q_1 and add the value obtained in step 3 to Q_3.

Step 5 Check the data set for any data value that is smaller than $Q_1 - 1.5(IQR)$ or larger than $Q_3 + 1.5(IQR)$.

This procedure is shown in Example 3–36.

EXAMPLE 3–36 Outliers

Check the following data set for outliers.

$$5, 6, 12, 13, 15, 18, 22, 50$$

SOLUTION

The data value 50 is extremely suspect. These are the steps in checking for an outlier.

Step 1 Find Q_1 and Q_3. $Q_1 = \frac{(6 + 12)}{2} = 9$; $Q_3 = \frac{(18 + 22)}{2} = 20$.

Step 2 Find the interquartile range (IQR), which is $Q_3 - Q_1$.

$$IQR = Q_3 - Q_1 = 20 - 9 = 11$$

Step 3 Multiply this value by 1.5.

$$1.5(11) = 16.5$$

Step 4 Subtract the value obtained in step 3 from Q_1, and add the value obtained in step 3 to Q_3.

$$9 - 16.5 = -7.5 \quad \text{and} \quad 20 + 16.5 = 36.5$$

Step 5 Check the data set for any data values that fall outside the interval from −7.5 to 36.5. The value 50 is outside this interval; hence, it can be considered an outlier.

There are several reasons why outliers may occur. First, the data value may have resulted from a measurement or observational error. Perhaps the researcher measured the variable incorrectly. Second, the data value may have resulted from a recording error. That is, it may have been written or typed incorrectly. Third, the data value may have been obtained from a subject that is not in the defined population. For example, suppose test scores were obtained from a seventh-grade class, but a student in that class was actually in the sixth grade and had special permission to attend the class. This student might have scored extremely low on that particular exam on that day. Fourth, the data value might be a legitimate value that occurred by chance (although the probability is extremely small).

There are no hard-and-fast rules on what to do with outliers, nor is there complete agreement among statisticians on ways to identify them. Obviously, if they occurred as a result of an error, an attempt should be made to correct the error or else the data value should be omitted entirely. When they occur naturally by chance, the statistician must make a decision about whether to include them in the data set.

When a distribution is normal or bell-shaped, data values that are beyond 3 standard deviations of the mean can be considered suspected outliers.

Applying the Concepts 3–3

Determining Dosages

In an attempt to determine necessary dosages of a new drug (HDL) used to control sepsis, assume you administer varying amounts of HDL to 40 mice. You create four groups and label them *low dosage*, *moderate dosage*, *large dosage*, and *very large dosage*. The dosages also vary within each group. After the mice are injected with the HDL and the sepsis bacteria, the time until the onset of sepsis is recorded. Your job as a statistician is to effectively communicate the results of the study.

1. Which measures of position could be used to help describe the data results?
2. If 40% of the mice in the top quartile survived after the injection, how many mice would that be?
3. What information can be given from using percentiles?
4. What information can be given from using quartiles?
5. What information can be given from using standard scores?

See page 184 for the answers.

Exercises 3–3

1. What is a *z* score?
2. Define *percentile rank*.
3. What is the difference between a percentage and a percentile?
4. Define *quartile*.
5. What is the relationship between quartiles and percentiles?
6. What is a decile?
7. How are deciles related to percentiles?
8. To which percentile, quartile, and decile does the median correspond?
9. **Vacation Days** If the average number of vacation days for a selection of various countries has a mean of 29.4 days and a standard deviation of 8.6 days, find the *z* scores for the average number of vacation days in each of these countries.

Canada	26 days
Italy	42 days
United States	13 days

Source: www.infoplease.com

10. **Age of Senators** The average age of Senators in the 114th congress was 61.7 years. If the standard deviation was 10.6, find the *z* scores of a senator who is 48 years old and one who is 66 years old.
11. **Marriage Age for Females** The mean age at which females marry is 24.6. The standard deviation is 3.2 years. Find the corresponding *z* score for each.
    a. 27 *d.* 18
    b. 22 *e.* 26
    c. 31
12. **Teacher's Salary** The average teacher's salary in a particular state is $54,166. If the standard deviation is $10,200, find the salaries corresponding to the following *z* scores.
    a. 2 *d.* 2.5
    b. −1 *e.* −1.6
    c. 0
13. **Test Scores** Which is a better relative position, a score of 83 on a geography test that has a mean of 72 and a standard deviation of 6, or a score of 61 on an accounting test that has a mean of 55 and a standard deviation of 3.5?
14. **College and University Debt** A student graduated from a 4-year college with an outstanding loan of $9650 where the average debt is $8455 with a standard deviation of $1865. Another student graduated from a university with an outstanding loan of $12,360 where the average of the outstanding loans was $10,326 with a standard deviation of $2143. Which student had a higher debt in relationship to his or her peers?

15. Annual Miles Driven The average miles driven annually per licensed driver in the United States is approximately 14,090 miles. If we assume a fairly mound-shaped distribution with a standard deviation of approximately 3500 miles, find the following.

a. z score for 16,000 miles

b. z score for 10,000 miles

c. Number of miles corresponding to z scores of 1.6, −0.5, and 0.

Source: World Almanac 2012

16. Which score indicates the highest relative position?

a. A score of 3.2 on a test with $\bar{X} = 4.6$ and $s = 1.5$

b. A score of 630 on a test with $\bar{X} = 800$ and $s = 200$

c. A score of 43 on a test with $\bar{X} = 50$ and $s = 5$

17. The data show the population (in thousands) for a recent year of a sample of cities in South Carolina.

29	26	15	13	17	58
14	25	37	19	40	67
23	10	97	12	129	
27	20	18	120	35	
66	21	11	43	22	

Source: U.S. Census Bureau

Find the data value that corresponds to each percentile.

a. 40th percentile

b. 75th percentile

c. 90th percentile

d. 30th percentile

Using the same data, find the percentile corresponding to the given data value.

e. 27

f. 40

g. 58

h. 67

18. College Room and Board Costs Room and board costs for selected schools are summarized in this distribution. Find the approximate cost of room and board corresponding to each of the following percentiles by constructing a percentile graph.

Costs (in dollars)	Frequency
3000.5–4000.5	5
4000.5–5000.5	6
5000.5–6000.5	18
6000.5–7000.5	24
7000.5–8000.5	19
8000.5–9000.5	8
9000.5–10,000.5	5

a. 30th percentile

b. 50th percentile

c. 75th percentile

d. 90th percentile

Source: World Almanac

Using the same data, find the approximate percentile rank of each of the following costs.

e. 5500 *g.* 6500

f. 7200 *h.* 8300

19. Achievement Test Scores The data shown represent the scores on a national achievement test for a group of 10th-grade students. Find the approximate percentile ranks of these scores by constructing a percentile graph.

a. 220 *d.* 280

b. 245 *e.* 300

c. 276

Score	Frequency
196.5–217.5	5
217.5–238.5	17
238.5–259.5	22
259.5–280.5	48
280.5–301.5	22
301.5–322.5	6

For the same data, find the approximate scores that correspond to these percentiles.

f. 15th *i.* 65th

g. 29th *j.* 80th

h. 43rd

20. Airplane Speeds The airborne speeds in miles per hour of 21 planes are shown. Find the approximate values that correspond to the given percentiles by constructing a percentile graph.

Class	Frequency
366–386	4
387–407	2
408–428	3
429–449	2
450–470	1
471–491	2
492–512	3
513–533	4
	21

Source: The World Almanac and Book of Facts

a. 9th *d.* 60th

b. 20th *e.* 75th

c. 45th

Using the same data, find the approximate percentile ranks of the following speeds in miles per hour (mph).

f. 380 mph *i.* 505 mph

g. 425 mph *j.* 525 mph

h. 455 mph

21. Average Weekly Earnings The average weekly earnings in dollars for various industries are listed below. Find the percentile rank of each value.

804 736 659 489 777 623 597 524 228

For the same data, what value corresponds to the 40th percentile?

Source: New York Times Almanac

22. Test Scores Find the percentile rank for each test score in the data set.

12, 28, 35, 42, 47, 49, 50

What value corresponds to the 60th percentile?

23. Hurricane Damage Find the percentile rank for each value in the data set. The data represent the values in billions of dollars of the damage of 10 hurricanes.

1.1, 1.7, 1.9, 2.1, 2.2, 2.5, 3.3, 6.2, 6.8, 20.3

What value corresponds to the 40th percentile?

Source: Insurance Services Office.

24. Test Scores Find the percentile rank for each test score in the data set.

5, 12, 15, 16, 20, 21

What test score corresponds to the 33rd percentile?

25. Taxes The data for a recent year show the taxes (in millions of dollars) received from a random sample of 10 states. Find the first and third quartiles and the IQR.

13 15 32 36 11 24 6 25 11 71

Source: U.S. Census Bureau.

26. Medical Marijuana 2015 Sales Tax: The data show the amount of sales tax paid in Denver County, Colorado. Find the first and third quartiles for the data.

Month	Sales Tax	Month	Sales Tax
Jan	363,061	July	518,868
Feb	358,208	August	554,013
March	418,500	September	506,809
April	266,771	October	341,421
May	399,814	November	349,026
June	453,698	December	532,545

Source: Colorado Department of Revenue

27. Gold Reserves The data show the gold reserves for a recent year for 11 world countries. Find the first and third quartiles and the IQR. The data are in millions of troy ounces.

33.9 78.3 108.9 17.9 78.8 24.6 19.7 33.3 33.4 10.0 261.5

Source: International Monetary Fund.

28. Police Calls in Schools The number of incidents in which police were needed for a sample of 9 schools in Allegheny County is 7, 37, 3, 8, 48, 11, 6, 0, 10. Find the first and third quartiles for the data.

29. Check each data set for outliers.

a. 46, 28, 32, 21, 25, 29, 34, 19

b. 82, 100, 97, 93, 89, 90, 65, 94, 101

c. 527, 1007, 489, 371, 175

30. Check each data set for outliers.

a. 88, 72, 97, 84, 86, 85, 100

b. 145, 119, 122, 118, 125, 116

c. 14, 16, 27, 18, 13, 19, 36, 15, 20

Extending the Concepts

31. Another measure of the average is called the *midquartile;* it is the numerical value halfway between Q_1 and Q_3, and the formula is

$$\text{Midquartile} = \frac{Q_1 + Q_3}{2}$$

Using this formula and other formulas, find Q_1, Q_2, Q_3, the midquartile, and the interquartile range for each data set.

a. 5, 12, 16, 25, 32, 38

b. 53, 62, 78, 94, 96, 99, 103

32. An employment evaluation exam has a variance of 250. Two particular exams with raw scores of 142 and 165 have z scores of −0.5 and 0.955, respectively. Find the mean of the distribution.

33. A particular standardized test has scores that have a mound-shaped distribution with mean equal to 125 and standard deviation equal to 18. Tom had a raw score of 158, Dick scored at the 98th percentile, and Harry had a z score of 2.00. Arrange these three students in order of their scores from lowest to highest. Explain your reasoning.

d) Edit the last dialog box and type **S** for the variable. In the expression box, drag the mouse over the previous expression to highlight it.

e) Click the button in the keypad for parentheses. Type **SQRT** at the beginning of the line, upper- or lowercase will work. The expression should be SQRT((sumX2-sumX**2/n)/(n-1)).

f) Click [OK].

Display Results

g) Select **Data>Display Data**, then highlight all columns and constants in the list.

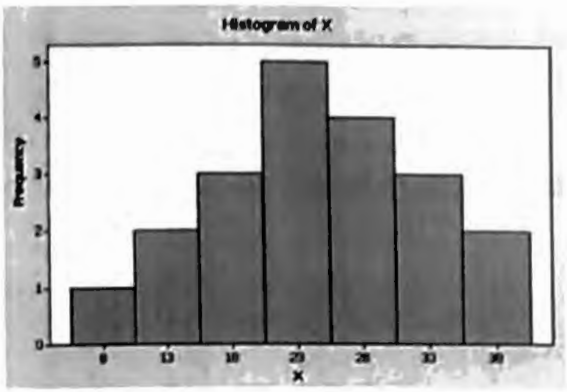
h) Click [Select] then [OK].

The session window will display all our work! Create the histogram with instructions from Chapter 2.

Data Display

n	20.0000
sumX	490.000
sumX2	13310.0

Row	X	f	fX	fX2	Mean	Variance	S
1	8	1	8	64	24.5	68.6842	8.28759
2	13	2	26	338			
3	18	3	54	972			
4	23	5	115	2645			
5	28	4	112	3136			
6	33	3	99	3267			
7	38	2	76	2888			



3–4 Exploratory Data Analysis

OBJECTIVE 4

Use the techniques of exploratory data analysis, including boxplots and five-number summaries, to discover various aspects of data

In traditional statistics, data are organized by using a frequency distribution. From this distribution various graphs such as the histogram, frequency polygon, and ogive can be constructed to determine the shape or nature of the distribution. In addition, various statistics such as the mean and standard deviation can be computed to summarize the data.

The purpose of traditional analysis is to confirm various conjectures about the nature of the data. For example, from a carefully designed study, a researcher might want to know if the proportion of Americans who are exercising today has increased from 10 years ago. This study would contain various assumptions about the population, various definitions such as the definition of exercise, and so on.

In **exploratory data analysis (EDA),** data can be organized using a *stem and leaf plot.* (See Chapter 2.) The measure of central tendency used in EDA is the *median.* The measure of variation used in EDA is the *interquartile range* Q_3–Q_1. In EDA the data are represented graphically using a *boxplot* (sometimes called a box and whisker plot). The purpose of exploratory data analysis is to examine data to find out what information can be discovered about the data, such as the center and the spread. Exploratory data analysis was developed by John Tukey and presented in his book *Exploratory Data Analysis* (Addison-Wesley, 1977).

The Five-Number Summary and Boxplots

A **boxplot** can be used to graphically represent the data set. These plots involve five specific values:

1. The lowest value of the data set (i.e., minimum)
2. Q_1
3. The median

4. Q_3
5. The highest value of the data set (i.e., maximum)

These values are called a **five-number summary** of the data set.

A boxplot is a graph of a data set obtained by drawing a horizontal line from the minimum data value to Q_1, drawing a horizontal line from Q_3 to the maximum data value, and drawing a box whose vertical sides pass through Q_1 and Q_3 with a vertical line inside the box passing through the median or Q_2.

Procedure Table

Constructing a Boxplot

Step 1	Find the five-number summary for the data.
Step 2	Draw a horizontal axis and place the scale on the axis. The scale should start on or below the minimum data value and end on or above the maximum data value.
Step 3	Locate the lowest data value, Q_1, the median, Q_3, and the highest data value; then draw a box whose vertical sides go through Q_1 and Q_3. Draw a vertical line through the median. Finally, draw a line from the minimum data value to the left side of the box, and draw a line from the maximum data value to the right side of the box.

EXAMPLE 3–37 Number of Meteorites Found

The number of meteorites found in 10 states of the United States is 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Construct a boxplot for the data.

Source: Natural History Museum.

SOLUTION

Step 1 Find the five-number summary for the data.

Arrange the data in order:

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

Find the median.

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

↑

Median

$$\text{Median} = \frac{78 + 89}{2} = 83.5$$

Find Q_1.

30, 39, 47, 48, 78

↑

Q_1

Find Q_3.

89, 138, 164, 215, 296

↑

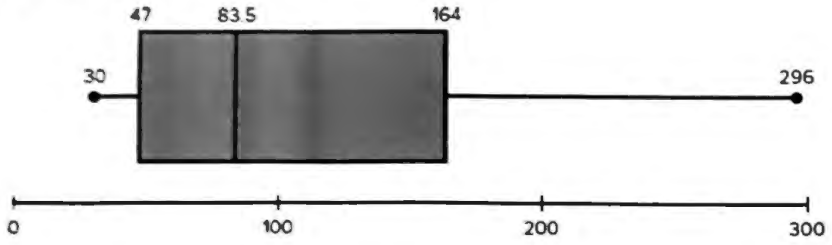
Q_3

The minimum data value is 30, and the maximum data value is 296.

Step 2 Draw a horizontal axis and the scale.

Step 3 Draw the box above the scale using Q_1 and Q_3. Draw a vertical line through the median, and draw lines from the lowest data value to the box and from the highest data value to the box. See Figure 3–7.

FIGURE 3–7 Boxplot for Example 3–37



Information Obtained from a Boxplot

1. *a.* If the median is near the center of the box, the distribution is approximately symmetric.
   b. If the median falls to the left of the center of the box, the distribution is positively skewed.
   c. If the median falls to the right of the center, the distribution is negatively skewed.
2. *a.* If the lines are about the same length, the distribution is approximately symmetric.
   b. If the right line is larger than the left line, the distribution is positively skewed.
   c. If the left line is larger than the right line, the distribution is negatively skewed.

The boxplot in Figure 3–7 indicates that the distribution is slightly positively skewed.

If the boxplots for two or more data sets are graphed on the same axis, the distributions can be compared. To compare the averages, use the location of the medians. To compare the variability, use the interquartile range, i.e., the length of the boxes. Example 3–38 shows this procedure.

EXAMPLE 3–38 Speeds of Roller Coasters

The data shown are the speeds in miles per hour of a sample of wooden roller coasters and a sample of steel roller coasters. Compare the distributions by using boxplots.

Wood				Steel			
50	56	60	48	55	70	48	28
35	67	72	68	100	106	102	120

Source: UltimateRollerCoaster.com

SOLUTION

Step 1 For the wooden coasters,

$$\begin{array}{cccccccc} 35 & 48 & 50 & 56 & 60 & 67 & 68 & 72 \\ & & \uparrow & & \uparrow & & \uparrow & \\ & & Q_1 & & \text{MD} & & Q_3 & \end{array}$$

$$Q_1 = \frac{48 + 50}{2} = 49 \qquad \text{MD} = \frac{56 + 60}{2} = 58 \qquad Q_3 = \frac{67 + 68}{2} = 67.5$$

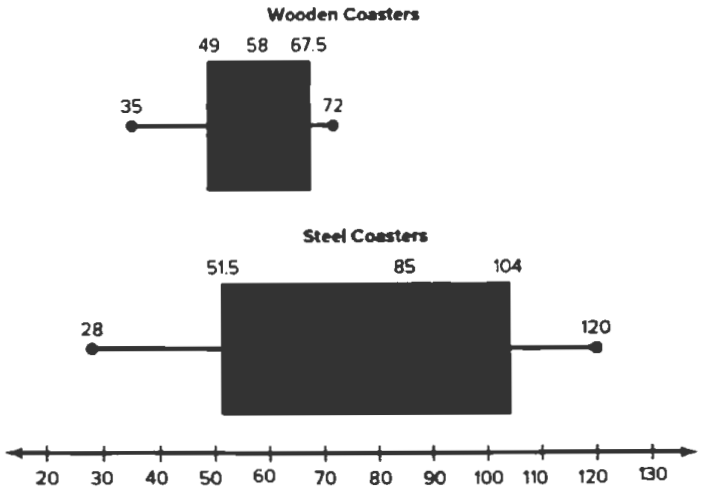
Step 2 For the steel coasters,

$$\begin{array}{cccccccc} 28 & 48 & 55 & 70 & 100 & 102 & 106 & 120 \\ & \uparrow & & \uparrow & & \uparrow & & \\ & Q_1 & & \text{MD} & & Q_3 & & \end{array}$$

$$Q_1 = \frac{48 + 55}{2} = 51.5 \qquad \text{MD} = \frac{70 + 100}{2} = 85 \qquad Q_3 = \frac{102 + 106}{2} = 104$$

Step 3 Draw the boxplots. See Figure 3–8.

FIGURE 3–8 Boxplots for Example 3–38



The boxplots show that the median of the speeds of the steel coasters is much higher than the median speeds of the wooden coasters. The interquartile range (spread) of the steel coasters is much larger than that of the wooden coasters. Finally, the range of the speeds of the steel coasters is larger than that of the wooden coasters.

A *modified boxplot* can be drawn and used to check for outliers. See Exercise 19 in Extending the Concepts in this section.

In exploratory data analysis, *hinges* are used instead of quartiles to construct boxplots. When the data set consists of an even number of values, hinges are the same as quartiles. Hinges for a data set with an odd number of values differ somewhat from quartiles. However, since most calculators and computer programs use quartiles, they will be used in this textbook.

Table 3–5 shows the correspondence between the traditional and the exploratory data analysis approach.

TABLE 3–5 Traditional versus EDA Techniques

Frequency distribution	Stem and leaf plot
Histogram	Boxplot
Mean	Median
Standard deviation	Interquartile range

Applying the Concepts 3–4

The Noisy Workplace

Assume you work for OSHA (Occupational Safety and Health Administration) and have complaints about noise levels from some of the workers at a state power plant. You charge the power plant with taking decibel readings at six different areas of the plant at different times of the day and week. The results of the data collection are listed. Use boxplots to initially explore the data and make recommendations about which plant areas workers must be provided with protective ear wear. The safe hearing level is approximately 120 decibels.

30	64	100	25	59	67
12	99	59	15	63	80
35	87	78	30	81	99
65	59	97	20	110	49
24	23	84	61	65	67
59	16	64	56	112	56
68	94	53	34	132	80
57	78	59	22	145	125
100	57	89	24	163	100
61	32	88	21	120	93
32	52	94	32	84	56
45	78	66	52	99	45
92	59	57	14	105	80
56	55	62	10	68	34
44	55	64	33	75	21

See page 184 for the answers.

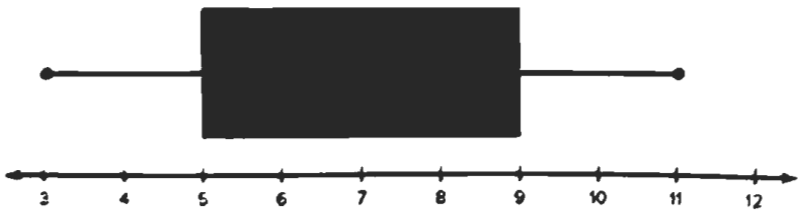
Exercises 3–4

For Exercises 1–6, identify the five-number summary and find the interquartile range.

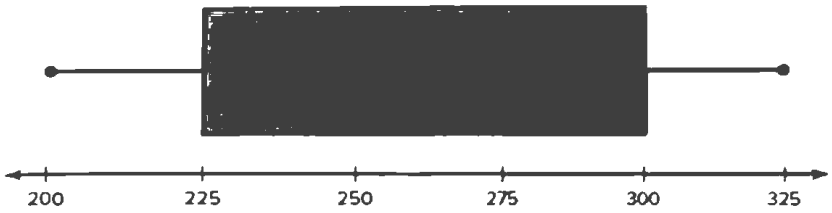
1. 8, 12, 32, 6, 27, 19, 54
2. 19, 16, 48, 22, 7
3. 362, 589, 437, 316, 192, 188
4. 147, 243, 156, 632, 543, 303
5. 14.6, 19.8, 16.3, 15.5, 18.2
6. 9.7, 4.6, 2.2, 3.7, 6.2, 9.4, 3.8

For Exercises 7–10, use each boxplot to identify the maximum value, minimum value, median, first quartile, third quartile, and interquartile range.

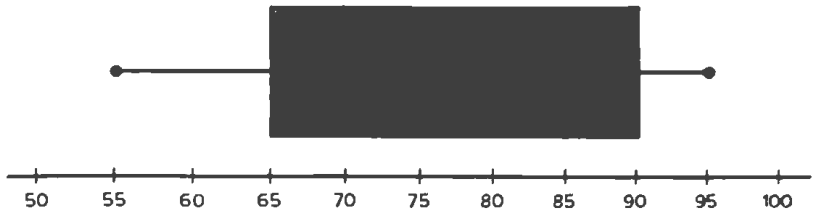
7.



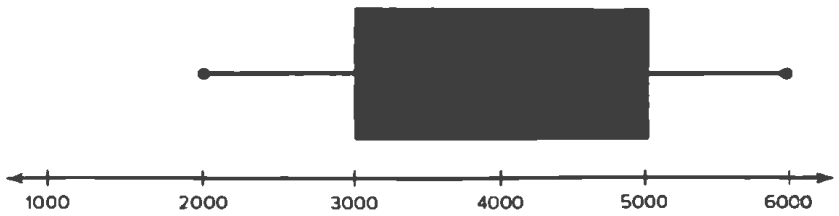
8.



9.



10.



11. **School Graduation Rates** The data show a sample of states' percentage of public high school graduation rates for a recent year. Construct a boxplot for the data, and comment on the shape of the distribution.

79 82 77 84 80 89 60 79 91 93 88

Source: U.S. Department of Education

12. **Innings Pitched** Construct a boxplot for the following data which represent the number of innings pitched by the ERA leaders for the past few years. Comment on the shape of the distribution.

239 266 245 236 241 246 240
249 251 238 228 248 232

Source: Baseball-Reference.com

13. **Population of Colonies** The data show the population (in thousands) of the U.S. Colonies in 1700 (Vermont was not a colony until 1791). Construct a boxplot and decide if the distribution is symmetric.

26.0 2.5 29.6 55.9 5.0 14.0 19.1 10.7
18.0 5.9 5.7 58.6

Source: U.S. Census Bureau

14. **Visitors Who Travel to Foreign Countries** Construct a boxplot for the number (in millions) of visitors who traveled to a foreign country each year for a random selection of years. Comment on the skewness of the distribution.

4.3 0.5 0.6 0.8 0.5
0.4 3.8 1.3 0.4 0.3

15. **Areas of Islands** The data show the sizes in square miles of notable islands in the Baltic Sea and the Aleutian Islands. Construct a boxplot for each data set and compare the distributions.

Baltic Sea	Aleutian Islands
610	275
228	1051
1154	571
1159	686
2772	350

16. **Size of Dams** These data represent the volumes in cubic yards of the largest dams in the United States and in South America. Construct a boxplot of the data for each region and compare the distributions.

United States	South America
125,628	311,539
92,000	274,026
78,008	105,944
77,700	102,014
66,500	56,242
62,850	46,563
52,435	
50,000	

Source: New York Times Almanac

17. **Largest Dams** The data show the heights (in feet) of the 10 largest dams in the United States. Identify the five-number summary and the interquartile range, and draw a boxplot.

770 730 717 710 645 606 602 585 578 564

Source: U.S. Army Corps of Engineers

positively skewed or right-skewed distribution 121
quartile 155
range 129
range rule of thumb 138
resistant statistic 157
sample mean 112
standard deviation 134
statistic 111
symmetric distribution 121
unimodal 116
variance 134
weighted mean 119
z score or standard score 148

Important Formulas

Formula for the mean for individual data:

Sample $\bar{X} = \frac{\Sigma X}{n}$ Population $\mu = \frac{\Sigma X}{N}$

Formula for the mean for grouped data:

$$\bar{X} = \frac{\Sigma f \cdot X_m}{n}$$

Formula for the weighted mean:

$$\bar{X} = \frac{\Sigma wX}{\Sigma w}$$

Formula for the midrange:

$$\text{MR} = \frac{\text{lowest value} + \text{highest value}}{2}$$

Formula for the range:

$$R = \text{highest value} - \text{lowest value}$$

Formula for the variance for population data:

$$\sigma^2 = \frac{\Sigma (X - \mu)^2}{N}$$

Formula for the variance for sample data (shortcut formula for the unbiased estimator):

$$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}$$

Formula for the variance for grouped data:

$$s^2 = \frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)}$$

Formula for the standard deviation for population data:

$$\sigma = \sqrt{\frac{\Sigma (X - \mu)^2}{N}}$$

Formula for the standard deviation for sample data (shortcut formula):

$$s = \sqrt{\frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n-1)}}$$

Formula for the standard deviation for grouped data:

$$s = \sqrt{\frac{n(\Sigma f \cdot X_m^2) - (\Sigma f \cdot X_m)^2}{n(n-1)}}$$

Formula for the coefficient of variation:

$$\text{CVar} = \frac{s}{\bar{X}} \cdot 100 \quad \text{or} \quad \text{CVar} = \frac{\sigma}{\mu} \cdot 100$$

Range rule of thumb:

$$s \approx \frac{\text{range}}{4}$$

Expression for Chebyshev's theorem: The proportion of values from a data set that will fall within k standard deviations of the mean will be at least

$$1 - \frac{1}{k^2}$$

where k is a number greater than 1.

Formula for the z score (standard score):

Sample $z = \frac{X - \bar{X}}{s}$ or Population $z = \frac{X - \mu}{\sigma}$

Formula for the cumulative percentage:

$$\text{Cumulative } \% = \frac{\text{cumulative frequency}}{n} \cdot 100$$

Formula for the percentile rank of a value X:

$$\text{Percentile} = \frac{\text{number of values below } X + 0.5}{\text{total number of values}} \cdot 100$$

Formula for finding a value corresponding to a given percentile:

$$c = \frac{n \cdot p}{100}$$

Formula for interquartile range:

$$\text{IQR} = Q_3 - Q_1$$

Review Exercises

SECTION 3–1

1. **Bank Failures** The data show the number of bank failures over a 15-year period. Find the mean, median, midrange, and mode for the data.

14	24	51	92	157	148	30	3
0	0	4	3	11	4	7	

Source: Federal Deposit Insurance Corporation

2. **Shark Attacks** The number of shark attacks and deaths over a recent 5-year period is shown. Find the mean, median, mode, and midrange for the data.

Attacks	71	64	61	65	57
Deaths	1	4	4	7	4

3. **Systolic Blood Pressure** The data show the systolic blood pressure of 30 college students. Find the mean and modal class.

Class	Frequency
105–109	2
110–114	5
115–119	6
120–124	8
125–129	8
130–134	1

4. **SAT Scores** The mean SAT math scores for selected states are represented. Find the mean class and modal class.

Score	Frequency
478–504	4
505–531	6
532–558	2
559–585	2
586–612	2

Source: World Almanac.

5. **Households of Four Television Networks** A survey showed the number of viewers and number of households of four television networks. Find the average number of viewers, using the weighted mean.

Households	1.4	0.8	0.3	1.6
Viewers (in millions)	1.6	0.8	0.4	1.8

Source: Nielsen Media Research.

6. **Investment Earnings** An investor calculated these percentages of each of three stock investments with payoffs as shown. Find the average payoff. Use the weighted mean.

Stock	Percent	Payoff
A	30	$10,000
B	50	3,000
C	20	1,000

SECTION 3–2

7. **Confirmed Measles Cases** The data show a sample of the number of confirmed measles cases over a recent 12-year period. Find the range, variance, and standard deviation for the data.

212 63 71 140 43 55 66 37 56 44 116 86

Source: Centers for Disease Control and Prevention.

8. **Tallest Buildings** The number of stories in the 13 tallest buildings in Houston are shown. Find the range, variance, and standard deviation for the data.

75 71 64 56 53 55 47 55 52 50 50 50 47

Source: World Almanac

9. **Rise in Tides** Shown here is a frequency distribution for the rise in tides at 30 selected locations in the United States. Find the variance and standard deviation for the data.

Rise in tides (inches)	Frequency
12.5–27.5	6
27.5–42.5	3
42.5–57.5	5
57.5–72.5	8
72.5–87.5	6
87.5–102.5	2

10. **Fuel Capacity** The fuel capacity in gallons for randomly selected cars is shown here. Find the variance and standard deviation for the data.

Class	Frequency
10–12	6
13–15	4
16–18	14
19–21	15
22–24	8
25–27	2
28–30	1
	50

11. If the range of a data set is 24, find the approximate value of the standard deviation, using the range rule of thumb.

12. If the range of a data set is 56, find the approximate value of the standard deviation, using the range rule of thumb.

13. **Textbooks in Professors' Offices** If the average number of textbooks in professors' offices is 16, the standard deviation is 5, and the average age of the professors is 43, with a standard deviation of 8, which data set is more variable?

14. **Magazines in Bookstores** A survey of bookstores showed that the average number of magazines carried is 56, with a standard deviation of 12. The same survey showed that the average length of time each store had been in business was 6 years, with a standard deviation of 2.5 years. Which is more variable, the number of magazines or the number of years?

15. **Cost of Car Rentals** A survey of car rental agencies shows that the average cost of a car rental is $0.32 per mile. The standard deviation is $0.03. Using Chebyshev's theorem, find the range in which at least 75% of the data values will fall.

16. Average Earnings of Workers The average earnings of year-round full-time workers 25–34 years old with a bachelor's degree or higher were $58,500 in 2003. If the standard deviation is $11,200, what can you say about the percentage of these workers who earn.

a. Between $47,300 and $69,700?
b. More than $80,900?
c. How likely is it that someone earns more than $100,000?

Source: New York Times Almanac

17. Labor Charges The average labor charge for automobile mechanics is $54 per hour. The standard deviation is $4. Find the minimum percentage of data values that will fall within the range of $48 to $60. Use Chebyshev's theorem.

18. Costs to Train Employees For a certain type of job, it costs a company an average of $231 to train an employee to perform the task. The standard deviation is $5. Find the minimum percentage of data values that will fall in the range of $219 to $243. Use Chebyshev's theorem.

19. Cost of a Man's Haircut The average cost of a man's haircut is $21. The standard deviation is $4. If the variable is approximately bell-shaped, within what limits would 68% of the haircut cost?

20. Exam Completion Time The mean time it takes a group of students to complete a statistics final exam is 44 minutes, and the standard deviation is 9 minutes. Within what limits would you expect approximately 95% of the students to complete the exam? Assume the variable is approximately normally distributed.

21. Cases of Meningitis The data show the number of specific recorded cases of meningitis for 14 specific states.

10 1 1 28 15 41 4 4 8 2 3
53 34 1

Source: Centers for Disease Control and Prevention

Find the z values for each

a. 10
b. 28
c. 41

SECTION 3–3

22. Exam Grades Which of these exam grades has a better relative position?

a. A grade of 82 on a test with $\bar{X} = 85$ and $s = 6$
b. A grade of 56 on a test with $\bar{X} = 60$ and $s = 5$

23. The number of police calls a small police department received each month is shown in the frequency distribution.

Class limits	Frequency
39.9–42.8	2
42.9–45.8	2
45.9–48.8	5
48.9–51.8	5
51.9–54.8	12
54.9–57.8	3

a. Construct a percentile graph.
b. Find the values that correspond to the 35th, 65th, and 85th percentiles.
c. Find the percentile of values 44, 48, and 54.

24. Printer Repairs The frequency distribution shows the number of days it took to fix each of 80 computer's printers.

Class limits	Frequency
1–3	7
4–6	9
7–9	32
10–12	20
13–15	12
	80

a. Construct a percentile graph.
b. Find the 20th, 50th, and 70th percentiles.
c. Find the percentile values of 5, 10, and 14.

25. Check each data set for outliers.

a. 506, 511, 517, 514, 400, 521
b. 3, 7, 9, 6, 8, 10, 14, 16, 20, 12

26. Check each data set for outliers.

a. 14, 18, 27, 26, 19, 13, 5, 25
b. 112, 157, 192, 116, 153, 129, 131

SECTION 3–4

27. Named Storms The data show the number of named storms for the years 1851–1860 and 1941 and 1950. Construct a boxplot for each data set and compare the distributions.

1851–1860	6	5	8	5	5	6	4	6	8	7
1941–1950	6	10	10	11	11	6	9	9	13	13

Source: National Hurricane Center

28. Hours Worked The data shown here represent the number of hours that 12 part-time employees at a toy store worked during the weeks before and after Christmas. Construct two boxplots and compare the distributions.

Before	38	16	18	24	12	30	35	32	31	30	24	35
After	26	15	12	18	24	32	14	18	16	18	22	12

4

Probability and Counting Rules

© Royalty-Free/CORBIS

STATISTICS TODAY

Would You Bet Your Life?

Humans not only bet money when they gamble, but also bet their lives by engaging in unhealthy activities such as smoking, drinking, using drugs, and exceeding the speed limit when driving. Many people don't care about the risks involved in these activities since they do not understand the concepts of probability. On the other hand, people may fear activities that involve little risk to health or life because these activities have been sensationalized by the press and media.

In his book *Probabilities in Everyday Life* (Ivy Books, p. 191), John D. McGervey states

> *When people have been asked to estimate the frequency of death from various causes, the most overestimated categories are those involving pregnancy, tornadoes, floods, fire, and homicide. The most underestimated categories include deaths from diseases such as diabetes, strokes, tuberculosis, asthma, and stomach cancer (although cancer in general is overestimated).*

The question then is, Would you feel safer if you flew across the United States on a commercial airline or if you drove? How much greater is the risk of one way to travel over the other? See Statistics Today—Revisited at the end of the chapter for the answer.

In this chapter, you will learn about probability—its meaning, how it is computed, and how to evaluate it in terms of the likelihood of an event actually happening.

OUTLINE

Introduction

4–1 Sample Spaces and Probability

4–2 The Addition Rules for Probability

4–3 The Multiplication Rules and Conditional Probability

4–4 Counting Rules

4–5 Probability and Counting Rules

Summary

OBJECTIVES

After completing this chapter, you should be able to:

1. Determine sample spaces and find the probability of an event, using classical probability or empirical probability.
2. Find the probability of compound events, using the addition rules.
3. Find the probability of compound events, using the multiplication rules.
4. Find the conditional probability of an event.
5. Find the total number of outcomes in a sequence of events, using the fundamental counting rule.
6. Find the number of ways that r objects can be selected from n objects, using the permutation rule.
7. Find the number of ways that r objects can be selected from n objects without regard to order, using the combination rule.
8. Find the probability of an event, using the counting rules.

Introduction

A cynical person once said, "The only two sure things are death and taxes." This philosophy no doubt arose because so much in people's lives is affected by chance. From the time you awake until you go to bed, you make decisions regarding the possible events that are governed at least in part by chance. For example, should you carry an umbrella to work today? Will your car battery last until spring? Should you accept that new job?

Probability as a general concept can be defined as the chance of an event occurring. Many people are familiar with probability from observing or playing games of chance, such as card games, slot machines, or lotteries. In addition to being used in games of chance, probability theory is used in the fields of insurance, investments, and weather forecasting and in various other areas. Finally, as stated in Chapter 1, probability is the basis of inferential statistics. For example, predictions are based on probability, and hypotheses are tested by using probability.

The basic concepts of probability are explained in this chapter. These concepts include *probability experiments, sample spaces,* the *addition* and *multiplication rules,* and the *probabilities of complementary events.* Also in this chapter, you will learn the rule for counting, the differences between permutations and combinations, and how to figure out how many different combinations for specific situations exist. Finally, Section 4–5 explains how the counting rules and the probability rules can be used together to solve a wide variety of problems.

4–1 Sample Spaces and Probability

The theory of probability grew out of the study of various games of chance using coins, dice, and cards. Since these devices lend themselves well to the application of concepts of probability, they will be used in this chapter as examples. This section begins by explaining some basic concepts of probability. Then the types of probability and probability rules are discussed.

Basic Concepts

Processes such as flipping a coin, rolling a die, or drawing a card from a deck are called *probability experiments.*

OBJECTIVE 1

Determine sample spaces and find the probability of an event, using classical probability or empirical probability

A probability experiment is a chance process that leads to well-defined results called outcomes.

An outcome is the result of a single trial of a probability experiment.

A trial means flipping a coin once, rolling one die once, or the like. When a coin is tossed, there are two possible outcomes: head or tail. (*Note:* We exclude the possibility of a coin landing on its edge.) In the roll of a single die, there are six possible outcomes: 1, 2, 3, 4, 5, or 6. In any experiment, the set of all possible outcomes is called the *sample space.*

A sample space is the set of all possible outcomes of a probability experiment.

Some sample spaces for various probability experiments are shown here.

Experiment	Sample space
Toss one coin	Head, tail
Roll a die	1, 2, 3, 4, 5, 6
Answer a true/false question	True, false
Toss two coins	Head-head, tail-tail, head-tail, tail-head

It is important to realize that when two coins are tossed, there are *four* possible outcomes, as shown in the fourth experiment above. Both coins could fall heads up. Both coins could fall tails up. Coin 1 could fall heads up and coin 2 tails up. Or coin 1 could fall tails up and coin 2 heads up. Heads and tails will be abbreviated as H and T throughout this chapter.

EXAMPLE 4–1 Rolling Dice

Find the sample space for rolling two dice.

SOLUTION

Since each die can land in six different ways, and two dice are rolled, the sample space can be presented by a rectangular array, as shown in Figure 4–1. The sample space is the list of pairs of numbers in the chart.

FIGURE 4–1
Sample Space for Rolling Two Dice (Example 4–1)

	Die 2					
Die 1	1	2	3	4	5	6
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

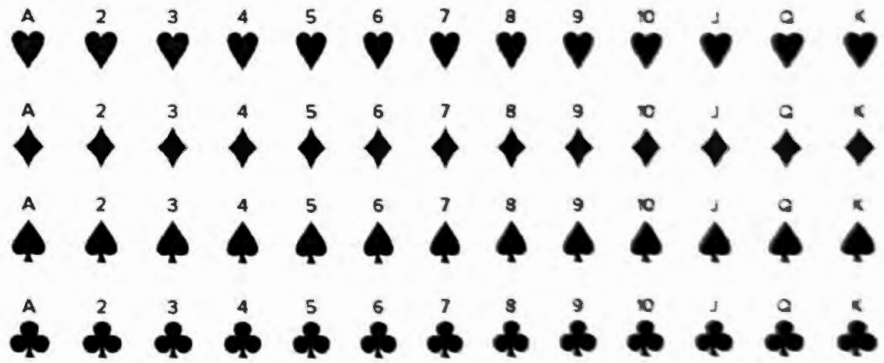
EXAMPLE 4–2 Drawing Cards

Find the sample space for drawing one card from an ordinary deck of cards.

SOLUTION

Since there are 4 suits (hearts, clubs, diamonds, and spades) and 13 cards for each suit (ace through king), there are 52 outcomes in the sample space. See Figure 4–2.

FIGURE 4–2 Sample Space for Drawing a Card (Example 4–2)



EXAMPLE 4–3 Gender of Children

Find the sample space for the gender of the children if a family has three children. Use B for boy and G for girl.

SOLUTION

There are two genders, boy and girl, and each child could be either gender. Hence, there are eight possibilities, as shown here.

BBB BBG BGB GBB GGG GGB GBG BGG

In Examples 4–1 through 4–3, the sample spaces were found by observation and reasoning; however, another way to find all possible outcomes of a probability experiment is to use a *tree diagram*.

A tree diagram is a device consisting of line segments emanating from a starting point and also from the outcome point. It is used to determine all possible outcomes of a probability experiment.

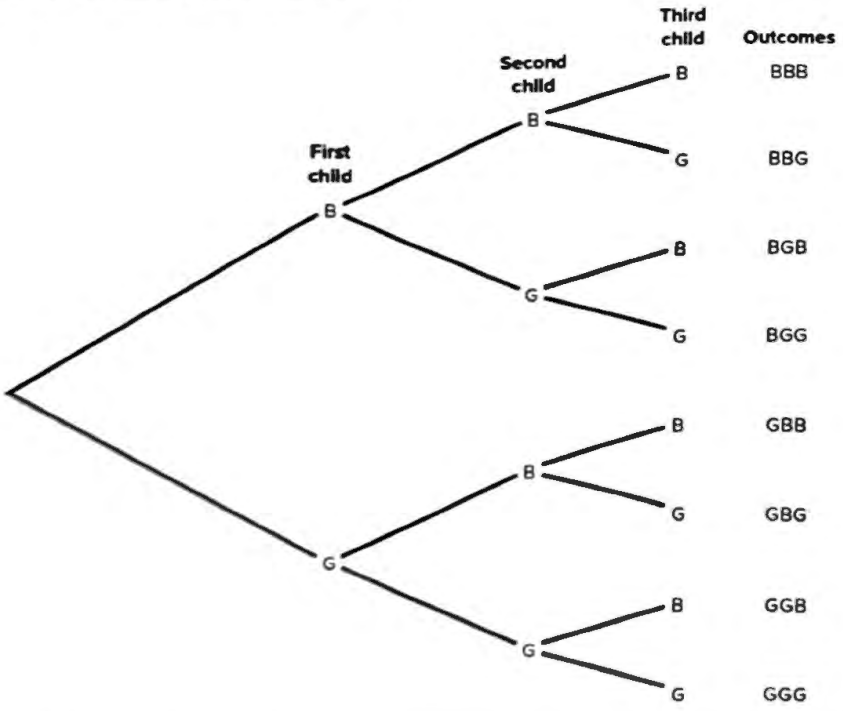
EXAMPLE 4–4 Gender of Children

Use a tree diagram to find the sample space for the gender of three children in a family, as in Example 4–3.

SOLUTION

Since there are two possibilities (boy or girl) for the first child, draw two branches from a starting point and label one B and the other G. Then if the first child is a boy, there are two possibilities for the second child (boy or girl), so draw two branches from B and label one B and the other G. Do the same if the first child is a girl. Follow the same procedure for the third child. The completed tree diagram is shown in Figure 4–3. To find the outcomes for the sample space, trace through all the possible branches, beginning at the starting point for each one.

FIGURE 4–3 Tree Diagram for Example 4–4



Historical Note

The famous Italian astronomer Galileo (1564–1642) found that sums of 10 and 11 occur more often than any other sum when three dice are tossed. Previously, it was thought that a sum of 9 occurred more often than any other sum.

Historical Note

A mathematician named Jerome Cardan (1501–1576) used his talents in mathematics and probability theory to make his living as a gambler. He is thought to be the first person to formulate the definition of classical probability.

An outcome was defined previously as the result of a single trial of a probability experiment. In many problems, one must find the probability of two or more outcomes. For this reason, it is necessary to distinguish between an outcome and an event.

An event consists of a set of outcomes of a probability experiment.

Historical Note

During the mid-1600s, a professional gambler named Chevalier de Méré made a considerable amount of money on a gambling game. He would bet unsuspecting patrons that in four rolls of a die, he could get at least one 6. He was so successful at the game that some people refused to play. He decided that a new game was necessary to continue his winnings. By reasoning, he figured he could roll at least one double 6 in 24 rolls of two dice, but his reasoning was incorrect and he lost systematically. Unable to figure out why, he contacted a mathematician named Blaise Pascal (1623–1662) to find out why.

Pascal became interested and began studying probability theory. He corresponded with a French government official, Pierre de Fermat (1601–1665), whose hobby was mathematics. Together the two formulated the beginnings of probability theory.

An event can be one outcome or more than one outcome. For example, if a die is rolled and a 6 shows, this result is called an *outcome*, since it is a result of a single trial. An event with one outcome is called a **simple event.** The event of getting an odd number when a die is rolled is called a **compound event,** since it consists of three outcomes or three simple events. In general, a compound event consists of two or more outcomes or simple events.

There are three basic interpretations of probability:

1. Classical probability
2. Empirical or relative frequency probability
3. Subjective probability

Classical Probability

Classical probability uses sample spaces to determine the numerical probability that an event will happen. You do not actually have to perform the experiment to determine that probability. Classical probability is so named because it was the first type of probability studied formally by mathematicians in the 17th and 18th centuries.

Classical probability assumes that all outcomes in the sample space are equally likely to occur. For example, when a single die is rolled, each outcome has the same probability of occurring. Since there are six outcomes, each outcome has a probability of $\frac{1}{6}$. When a card is selected from an ordinary deck of 52 cards, you assume that the deck has been shuffled, and each card has the same probability of being selected. In this case, it is $\frac{1}{52}$.

Equally likely events are events that have the same probability of occurring.

Formula for Classical Probability

The probability of any event E is

$$\frac{\text{Number of outcomes in } E}{\text{Total number of outcomes in the sample space}}$$

This probability is denoted by

$$P(E) = \frac{n(E)}{n(S)}$$

where $n(E)$ is the number of outcomes in E and $n(S)$ is the number of outcomes in the sample space S.

Probabilities can be expressed as fractions, decimals, or—where appropriate—percentages. If you ask, "What is the probability of getting a head when a coin is tossed?" typical responses can be any of the following three.

"One-half."

"Point five."

"Fifty percent."[1]

These answers are all equivalent. In most cases, the answers to examples and exercises given in this chapter are expressed as fractions or decimals, but percentages are used where appropriate.

[1]Strictly speaking, a percent is not a probability. However, in everyday language, probabilities are often expressed as percents (i.e., there is a 60% chance of rain tomorrow). For this reason, some probabilities will be expressed as percents throughout this book.

Rounding Rule for Probabilities Probabilities should be expressed as reduced fractions or rounded to three decimal places. When the probability of an event is an extremely small decimal, it is permissible to round the decimal to the first nonzero digit after the point. For example, 0.0000587 would be 0.00006. When obtaining probabilities from one of the tables in Appendix A, use the number of decimal places given in the table. If decimals are converted to percentages to express probabilities, move the decimal point two places to the right and add a percent sign.

EXAMPLE 4–5 Drawing Cards

Find the probability of getting a black 6 when one card is randomly selected from an ordinary deck.

SOLUTION

There are 52 cards in an ordinary deck, and there are two black 6s, that is, the 6 of clubs and the 6 of spades. Hence, the probability of getting a black 6 is

$$\frac{2}{52} = \frac{1}{26} \approx 0.038$$

EXAMPLE 4–6 Gender of Children

If a family has three children, find the probability that exactly two of the three children are boys.

SOLUTION

The sample space for the gender of three children has eight outcomes BBB, BBG, BGB, GBB, GGG, GGB, GBG, and BGG. (See Examples 4–3 and 4–4.) Since there are three ways to get two boys and one girl, that is BBG, BGB, and GBB, the probability of having exactly two boys is $\frac{3}{8}$.

Historical Note

Ancient Greeks and Romans made crude dice from animal bones, various stones, minerals, and ivory. When the dice were tested mathematically, some were found to be quite accurate.

In probability theory, it is important to understand the meaning of the words *and* and *or*. For example, if you were asked to find the probability of getting a queen *and* a heart when you were drawing a single card from a deck, you would be looking for the queen of hearts. Here the word *and* means "at the same time." The word *or* has two meanings. For example, if you were asked to find the probability of selecting a queen *or* a heart when one card is selected from a deck, you would be looking for one of the 4 queens or one of the 13 hearts. In this case, the queen of hearts would be included in both cases and counted twice. So there would be $4 + 13 - 1 = 16$ possibilities.

On the other hand, if you were asked to find the probability of getting a queen *or* a king, you would be looking for one of the 4 queens or one of the 4 kings. In this case, there would be $4 + 4 = 8$ possibilities. In the first case, both events can occur at the same time; we say that this is an example of the *inclusive or*. In the second case, both events cannot occur at the same time, and we say that this is an example of the *exclusive or*.

EXAMPLE 4–7 Drawing Cards

A card is drawn from an ordinary deck. Find the probability of getting

a A heart

b A black card

c The 8 of diamonds

d A queen

e A face card

SOLUTION

a. Refer to the sample space shown in Figure 4–2. There are 13 hearts in a deck of 52 cards; hence,

$$P(\heartsuit) = \frac{13}{52} = \frac{1}{4} = 0.25$$

b. There are 26 black cards in a deck, that is, 13 clubs and 13 spades. So the probability is

$$P(\text{black card}) = \frac{26}{52} = \frac{1}{2} = 0.5$$

c. There is one 8 of diamonds in a deck of 52 cards, so the probability is

$$P(8\diamondsuit) = \frac{1}{52} \approx 0.019$$

d. There are four queens in a deck of 52 cards; hence,

$$P(\text{queen}) = \frac{4}{52} = \frac{1}{13} \approx 0.077$$

e. There are 12 face cards in an ordinary deck of cards, that is, 4 suits (diamonds, hearts, spades, and clubs) and 3 face cards of each suit (jack, queen, and king), so

$$P(\text{face card}) = \frac{12}{52} = \frac{3}{13} \approx 0.231$$

There are four basic probability rules. These rules are helpful in solving probability problems, in understanding the nature of probability, and in deciding if your answers to the problems are correct.

Probability Rules

1. The probability of any event E is a number (either a fraction or decimal) between and including 0 and 1. This is denoted by $0 \le P(E) \le 1$.
2. The sum of the probabilities of all the outcomes in a sample space is 1.
3. If an event E cannot occur (i.e., the event contains no members in the sample space), its probability is 0.
4. If an event E is certain, then the probability of E is 1.

Historical Note

Paintings in tombs excavated in Egypt show that the Egyptians played games of chance. One game called *Hounds and Jackals* played in 1800 B.C. is similar to the present-day game of *Snakes and Ladders*.

Rule 1 states that probability values range from 0 to 1. When the probability of an event is close to 0, its occurrence is highly unlikely. When the probability of an event is near 0.5, there is about a 50-50 chance that the event will occur; and when the probability of an event is close to 1, the event is highly likely to occur. See Figure 4–4.

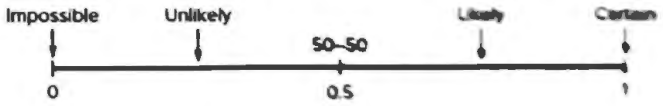


FIGURE 4–4
Range of Probability

Rule 2 can be illustrated by the example of rolling a single die. Each outcome in the sample space has a probability of $\frac{1}{6}$, and the sum of the probabilities of all the outcomes is 1, as shown.

Outcome	1		2		3		4		5		6
Probability	$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$		$\frac{1}{6}$
Sum	$\frac{1}{6}$	+	$\frac{1}{6}$	+	$\frac{1}{6}$	+	$\frac{1}{6}$	+	$\frac{1}{6}$	+	$\frac{1}{6} = \frac{6}{6} = 1$

Rule 3 is illustrated in Example 4–8.

EXAMPLE 4–8 Rolling a Die

When a single die is rolled, find the probability of getting a 9.

SOLUTION

Since the sample space is 1, 2, 3, 4, 5, and 6, it is impossible to get a 9. Hence, the probability is $P(9) = \frac{0}{6} = 0$.

Rule 4 states that if $P(E) = 1$, then the event E is certain to occur. This rule is illustrated in Example 4–9.

EXAMPLE 4–9 Rolling a Die

When a single die is rolled, what is the probability of getting a number less than 7?

SOLUTION

Since all outcomes—1, 2, 3, 4, 5, and 6—are less than 7, the probability is

$$P(\text{number less than } 7) = \frac{6}{6} = 1$$

The event of getting a number less than 7 is certain.

Complementary Events

Another important concept in probability theory is that of *complementary events*. When a die is rolled, for instance, the sample space consists of the outcomes 1, 2, 3, 4, 5, and 6. The event E of getting odd numbers consists of the outcomes 1, 3, and 5. The event of not getting an odd number is called the *complement* of event E, and it consists of the outcomes 2, 4, and 6.

The complement of an event E is the set of outcomes in the sample space that are not included in the outcomes of event E. The complement of E is denoted by $\overline{E}$ (read "E bar").

Example 4–10 further illustrates the concept of complementary events.

EXAMPLE 4–10 Finding Complements

Find the complement of each event:

a. Selecting a month that has 30 days

b. Selecting a day of the week that begins with the letter S

c. Rolling two dice and getting a number whose sum is 7

d. Selecting a letter of the alphabet (excluding y) that is a vowel

SOLUTION

a. Selecting a month that has 28 or 31 days, that is, January, February, March, May, July, August, October, or December

b. Selecting a day of the week that does not begin with S, that is, Monday, Tuesday, Wednesday, Thursday, or Friday

c. Rolling two dice and getting a sum of 2, 3, 4, 5, 6, 8, 9, 10, 11, or 12

d. Selecting a letter of the alphabet that is a consonant

The outcomes of an event and the outcomes of the complement make up the entire sample space. For example, if two coins are tossed, the sample space is HH, HT, TH, and TT. The complement of "getting all heads" is not "getting all tails," since the event "all heads" is HH, and the complement of HH is HT, TH, and TT. Hence, the complement of the event "all heads" is the event "getting at least one tail."

Since the event and its complement make up the entire sample space, it follows that the sum of the probability of the event and the probability of its complement will equal 1. That is, $P(E) + P(\overline{E}) = 1$. For example, let E = all heads, or HH, and let $\overline{E}$ = at least one tail, or HT, TH, TT. Then $P(E) = \frac{1}{4}$ and $P(\overline{E}) = \frac{3}{4}$; hence, $P(E) + P(\overline{E}) = \frac{1}{4} + \frac{3}{4} = 1$.

The rule for complementary events can be stated algebraically in three ways.

Rule for Complementary Events

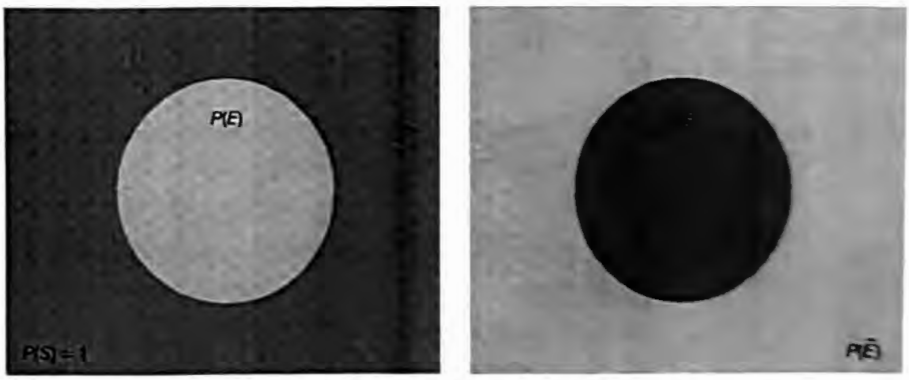
$$P(\overline{E}) = 1 - P(E) \quad \text{or} \quad P(E) = 1 - P(\overline{E}) \quad \text{or} \quad P(E) + P(\overline{E}) = 1$$

Stated in words, the rule is: *If the probability of an event or the probability of its complement is known, then the other can be found by subtracting the probability from* 1. This rule is important in probability theory because at times the best solution to a problem is to find the probability of the complement of an event and then subtract from 1 to get the probability of the event itself.

Probabilities can be represented pictorially by **Venn diagrams.** Figure 4–5(a) shows the probability of a simple event E. The area inside the circle represents the probability of event E, that is, $P(E)$. The area inside the rectangle represents the probability of all the events in the sample space $P(S)$.

The Venn diagram that represents the probability of the complement of an event $P(\overline{E})$ is shown in Figure 4–5(b). In this case, $P(\overline{E}) = 1 - P(E)$, which is the area inside the rectangle but outside the circle representing $P(E)$. Recall that $P(S) = 1$ and $P(E) = 1 - P(\overline{E})$. The reasoning is that $P(E)$ is represented by the area of the circle and $P(\overline{E})$ is the probability of the events that are outside the circle.

FIGURE 4–5
Venn Diagram for the Probability and Complement



(a) Simple probability (b) $P(\overline{E}) = 1 - P(E)$

EXAMPLE 4–11 Victims of Violence

In a study, it was found that 24% of people who were victims of a violent crime were age 20 to 24. If a person is selected at random, find the probability that the person is younger than 20 or older than 24.

Source: Based on statistics from the BFI.

SOLUTION

$$P(\text{not aged 20 to 24}) = 1 - P(\text{aged 20 to 24})$$
$$= 1 - 0.24 = 0.76 = 76\%$$

Empirical Probability

The difference between classical and **empirical probability** is that classical probability assumes that certain outcomes are equally likely (such as the outcomes when a die is rolled), while empirical probability relies on actual experience to determine the likelihood of outcomes. In empirical probability, one might actually roll a given die 6000 times, observe the various frequencies, and use these frequencies to determine the probability of an outcome.

Suppose, for example, that a researcher for the American Automobile Association (AAA) asked 50 people who plan to travel over the Thanksgiving holiday how they will get to their destination. The results can be categorized in a frequency distribution as shown.

Method	Frequency
Drive	41
Fly	6
Train or bus	3
	50

Now probabilities can be computed for various categories. For example, the probability of selecting a person who is driving is $\frac{41}{50}$, since 41 out of the 50 people said that they were driving.

Formula for Empirical Probability

Given a frequency distribution, the probability of an event being in a given class is

$$P(E) = \frac{\text{frequency for the class}}{\text{total frequencies in the distribution}} = \frac{f}{n}$$

This probability is called *empirical probability* and is based on observation.

EXAMPLE 4–12 Travel Survey

In the travel survey just described, find the probability that the person will travel by train or bus over the Thanksgiving holiday.

SOLUTION

$$P(E) = \frac{f}{n} = \frac{3}{50} = 0.06$$

Note: These figures are based on an AAA survey.

EXAMPLE 4–13 Distribution of Blood Types

In a sample of 50 people, 21 had type O blood, 22 had type A blood, 5 had type B blood, and 2 had type AB blood. Set up a frequency distribution and find the following probabilities.

a. A person has type O blood.

b. A person has type A or type B blood.

c. A person has neither type A nor type O blood.

d. A person does not have type AB blood.

Source: The American Red Cross

SOLUTION

Type	Frequency
A	22
B	5
AB	2
O	21
	Total 50

a. $P(\text{O}) = \frac{f}{n} = \frac{21}{50}$

b. $P(\text{A or B}) = \frac{22}{50} + \frac{5}{50} = \frac{27}{50}$

(Add the frequencies of the two classes.)

c. $P(\text{neither A nor O}) = \frac{5}{50} + \frac{2}{50} = \frac{7}{50}$

(Neither A nor O means that a person has either type B or type AB blood.)

d. $P(\text{not AB}) = 1 - P(\text{AB}) = 1 - \frac{2}{50} = \frac{48}{50} = \frac{24}{25}$

(Find the probability of not AB by subtracting the probability of type AB from 1.)

EXAMPLE 4–14 Sleep Hours

A recent survey found the following distribution for the number of hours a person sleeps per night.

Less than 6 hours	12
6 hours	26
7 hours	30
8 hours	28
More than 8 hours	4
	100

Find these probabilities for a person selected at random:

a. The person sleeps 8 hours per night.

b. The person sleeps fewer than 7 hours per night.

c. The person sleeps at most 8 hours per night.

d. The person sleeps at least 7 hours per night.

SOLUTION

a. $P(8 \text{ hours}) = \frac{28}{100} = \frac{7}{25}$

b. $P(\text{fewer than 7 hours}) = \frac{26}{100} + \frac{12}{100} = \frac{38}{100} = \frac{19}{50}$

c. $P(\text{at most 8 hours}) = P(\text{At most 8 hours means 8 or less hours})$

$= \frac{28}{100} + \frac{30}{100} + \frac{26}{100} + \frac{12}{100} = \frac{96}{100} = \frac{24}{25}$

d. $P(\text{at least 7 hours per night}) = P(\text{At least 7 hours per night means 7 or more hours per night}) = \frac{30}{100} + \frac{28}{100} + \frac{4}{100} = \frac{62}{100} = \frac{31}{50}$

Empirical probabilities can also be found by using a relative frequency distribution, as shown in Section 2–2. For example, the relative frequency distribution of the travel survey shown previously is

Method	Frequency	Relative frequency
Drive	41	0.82
Fly	6	0.12
Train or bus	3	0.06
	50	1.00

These frequencies are the same as the relative frequencies explained in Chapter 2.

Law of Large Numbers

When a coin is tossed one time, it is common knowledge that the probability of getting a head is $\frac{1}{2}$. But what happens when the coin is tossed 50 times? Will it come up heads 25 times? Not all the time. You should expect about 25 heads if the coin is fair. But due to chance variation, 25 heads will not occur most of the time.

If the empirical probability of getting a head is computed by using a small number of trials, it is usually not exactly $\frac{1}{2}$. However, as the number of trials increases, the empirical probability of getting a head will approach the theoretical probability of $\frac{1}{2}$, if in fact the coin is fair (i.e., balanced). This phenomenon is an example of the **law of large numbers.**

You should be careful to not think that the number of heads and number of tails tend to "even out." As the number of trials increases, the proportion of heads to the total number of trials will approach $\frac{1}{2}$. This law holds for any type of gambling game—tossing dice, playing roulette, and so on.

It should be pointed out that the probabilities that the proportions steadily approach may or may not agree with those theorized in the classical model. If not, it can have important implications, such as "the die is not fair." Pit bosses in Las Vegas watch for empirical trends that do not agree with classical theories, and they will sometimes take a set of dice out of play if observed frequencies are too far out of line with classical expected frequencies.

Subjective Probability

The third type of probability is called *subjective probability.* **Subjective probability** uses a probability value based on an educated guess or estimate, employing opinions and inexact information.

In subjective probability, a person or group makes an educated guess at the chance that an event will occur. This guess is based on the person's experience and evaluation of a solution. For example, a sportswriter may say that there is a 70% probability that the Pirates will win the pennant next year. A physician might say that, on the basis of her diagnosis, there is a 30% chance the patient will need an operation. A seismologist might say there is an 80% probability that an earthquake will occur in a certain area. These are only a few examples of how subjective probability is used in everyday life.

All three types of probability (classical, empirical, and subjective) are used to solve a variety of problems in business, engineering, and other fields.

Probability and Risk Taking

An area in which people fail to understand probability is risk taking. Actually, people fear situations or events that have a relatively small probability of happening rather than those events that have a greater likelihood of occurring. For example, many people think that the crime rate is increasing every year. However, in his book entitled *How Risk Affects Your Everyday Life,* author James Walsh states: "Despite widespread concern about the number of crimes committed in the United States, FBI and Justice Department statistics show that the national crime rate has remained fairly level for 20 years. It even dropped slightly in the early 1990s."

He further states, "Today most media coverage of risk to health and well-being focuses on shock and outrage." Shock and outrage make good stories and can scare us about the wrong dangers. For example, the author states that if a person is 20% overweight, the loss of life expectancy is 900 days (about 3 years), but loss of life expectancy from exposure to radiation emitted by nuclear power plants is 0.02 day. As you can see, being overweight is much more of a threat than being exposed to radioactive emission.

Many people gamble daily with their lives, for example, by using tobacco, drinking and driving, and riding motorcycles. When people are asked to estimate the probabilities or frequencies of death from various causes, they tend to overestimate causes such as accidents, fires, and floods and to underestimate the probabilities of death from diseases (other than cancer), strokes, etc. For example, most people think that their chances of dying of a heart attack are 1 in 20, when in fact they are almost 1 in 3; the chances of dying by pesticide poisoning are 1 in 200,000 (*True Odds* by James Walsh). The reason people think this way is that the news media sensationalize deaths resulting from catastrophic events and rarely mention deaths from disease.

When you are dealing with life-threatening catastrophes such as hurricanes, floods, automobile accidents, or texting while driving, it is important to get the facts. That is, get the actual numbers from accredited statistical agencies or reliable statistical studies, and then compute the probabilities and make decisions based on your knowledge of probability and statistics.

In summary, then, when you make a decision or plan a course of action based on probability, make sure that you understand the true probability of the event occurring. Also, find out how the information was obtained (i.e., from a reliable source). Weigh the cost of the action and decide if it is worth it. Finally, look for other alternatives or courses of action with less risk involved.

Applying the Concepts 4–1

Tossing a Coin

Assume you are at a carnival and decide to play one of the games. You spot a table where a person is flipping a coin, and since you have an understanding of basic probability, you believe that the odds of winning are in your favor. When you get to the table, you find out that all you have to do is to guess which side of the coin will be facing up after it is tossed. You are assured that the coin is fair, meaning that each of the two sides has an equally likely chance of occurring. You think back about what you learned in your statistics class about probability before you decide what to bet on. Answer the following questions about the coin-tossing game.

1. What is the sample space?
2. What are the possible outcomes?
3. What does the classical approach to probability say about computing probabilities for this type of problem?

You decide to bet on heads, believing that it has a 50% chance of coming up. A friend of yours, who had been playing the game for awhile before you got there, tells you that heads has come up the last 9 times in a row. You remember the law of large numbers.

4. What is the law of large numbers, and does it change your thoughts about what will occur on the next toss?
5. What does the empirical approach to probability say about this problem, and could you use it to solve this problem?
6. Can subjective probabilities be used to help solve this problem? Explain.
7. Assume you could win $1 million if you could guess what the results of the next toss will be. What would you bet on? Why?

See pages 253–255 for the answers.

Exercises 4–1

1. What is a probability experiment?

2. Define *sample space*.

3. What is the difference between an outcome and an event?

4. What are equally likely events?

5. What is the range of the values of the probability of an event?

6. When an event is certain to occur, what is its probability?

7. If an event cannot happen, what value is assigned to its probability?

8. What is the sum of the probabilities of all the outcomes in a sample space?

9. If the probability that it will rain tomorrow is 0.20, what is the probability that it won't rain tomorrow? Would you recommend taking an umbrella?

10. A probability experiment is conducted. Which of these cannot be considered a probability outcome?

a. $\frac{2}{5}$	*d.* 1.65	*g.* 1
b. 0.63	*e.* -0.44	*h.* 125%
c. $-\frac{1}{5}$	*f.* 0	*i.* 24%

11. Classify each statement as an example of classical probability, empirical probability, or subjective probability.
    a. The probability that a person will watch the 6 o'clock evening news is 0.15.
    b. The probability of winning at a Chuck-a-Luck game is $\frac{5}{36}$.
    c. The probability that a bus will be in an accident on a specific run is about 6%.
    d. The probability of getting a royal flush when five cards are selected at random is $\frac{1}{649,740}$.

12. Classify each statement as an example of classical probability, empirical probability, or subjective probability.
    a. The probability that a student will get a C or better in a statistics course is about 70%.
    b. The probability that a new fast-food restaurant will be a success in Chicago is 35%.
    c. The probability that interest rates will rise in the next 6 months is 0.50.
    d. The probability that the unemployment rate will fall next month is 0.03.

13. **Rolling a Die** If a die is rolled one time, find these probabilities:
    a. Getting a 7
    b. Getting an odd number
    c. Getting a number less than 7
    d. Getting a prime number (2, 3, or 5)

14. **Rolling a Die** If a die is rolled one time, find these probabilities:
    a. Getting a number less than 7.
    b. Getting a number greater than or equal to 3
    c. Getting a number greater than 2 and an even number
    d. Getting a number less than 1

15. **Rolling Two Dice** If two dice are rolled one time, find the probability of getting these results:
    a. A sum of 5
    b. A sum of 9 or 10
    c. Doubles

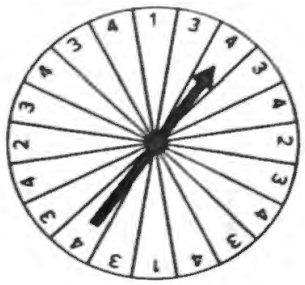
16. **Rolling Two Dice** If two dice are rolled one time, find the probability of getting these results:
    a. A sum less than 9
    b. A sum greater than or equal to 10
    c. A 3 on one die or on both dice.

17. **Drawing a Card** If one card is drawn from a deck, find the probability of getting these results:
    a. An ace
    b. A heart
    c. A 6 of spades
    d. A 10 or a jack
    e. A card whose face values less than 7 (Count aces as 1.)

18. **Drawing a Card** If a card is drawn from a deck, find the probability of getting these results:
    a. A 6 and a spade
    b. A black king
    c. A red card and a 7
    d. A diamond or a heart
    e. A black card

19. **Shopping Mall Promotion** A shopping mall has set up a promotion as follows. With any mall purchase of \$50 or more, the customer gets to spin the wheel shown here. If a number 1 comes up, the customer wins \$10. If the number 2 comes up, the customer wins \$5; and if the number 3 or 4 comes up, the customer wins a discount coupon. Find the following probabilities.
    a. The customer wins \$10.
    b. The customer wins money.

c. The customer wins a coupon.



20. Selecting a State Choose one of the 50 states at random.

a. What is the probability that it begins with the letter M?

b. What is the probability that it doesn't begin with a vowel?

21. Human Blood Types Human blood is grouped into four types. The percentages of Americans with each type are listed below.

O 43% A 40% B 12% AB 5%

Choose one American at random. Find the probability that this person

a. Has type B blood

b. Has type AB or O blood

c. Does not have type O blood

22. 2014 Top Albums (Based on U.S. sales) Of all of the U.S. album sales 1989 (Taylor Swift) accounted for 25% of sales, Frozen (Various Artists) accounted for 24.1% of sales, In the Lonely Hour (Sam Smith) accounted for 8.2% of sales. What is the probability that a randomly selected album was something other than these three albums?

Source: 2014 Nielsen Music U.S. Report.

23. Prime Numbers A prime number is a number that is evenly divisible only by 1 and itself. The prime numbers less than 100 are listed below.

2	3	5	7	11	13	17	19	23	29	31
37	41	43	47	53	59	61	67	71	73	79
83	89	97								

Choose one of these numbers at random. Find the probability that

a. The number is odd

b. The sum of the digits is odd

c. The number is greater than 70

24. Rural Speed Limits Rural speed limits for all 50 states are indicated below.

60 mph	65 mph	70 mph	75 mph
1 (HI)	18	18	13

Choose one state at random. Find the probability that its speed limit is

a. 60 or 70 miles per hour

b. Greater than 65 miles per hour

c. 70 miles per hour or less

Source: World Almanac

25. Gender of Children A couple has 4 children. Find each probability.

a. All girls

b. Exactly two girls and two boys

c. At least one child who is a girl

d. At least one child of each gender

26. Sources of Energy Uses in the United States A breakdown of the sources of energy used in the United States is shown below. Choose one energy source at random. Find the probability that it is

a. Not oil

b. Natural gas or oil

c. Nuclear

Oil 39%	Natural gas 24%	Coal 23%
Nuclear 8%	Hydropower 3%	Other 3%

Source: www.infoplease.com

27. Game of Craps In a game of craps, a player wins on the first roll if the player rolls a sum of 7 or 11, and the player loses if the player rolls a 2, 3, or 12. Find the probability that the game will last only one roll.

28. Computers in Elementary Schools Elementary and secondary schools were classified by the number of computers they had.

Computers	1–10	11–20	21–50	51–100	100+
Schools	3170	4590	16,741	23,753	34,803

Choose one school at random. Find the probability that it has

a. 50 or fewer computers

b. More than 100 computers

c. No more than 20 computers

Source: World Almanac

29. College Debt The following information shows the amount of debt students who graduated from college incur for a specific year.

$1 to $5000	$5001 to $20,000	$20,001 to $50,000	$50,000+
27%	40%	19%	14%

If a person who graduates has some debt, find the probability that

a. It is less than $5001

b. It is more than $20,000

c. It is between $1 and $20,000

d. It is more than $50,000

Source: USA TODAY

30. Population of Hawaii The population of Hawaii is 22.7% white, 1.5% African-American, 37.7% Asian, 0.2% Native American/Alaskan, 9.46% Native Hawaiian/Pacific Islander, 8.9% Hispanic, 19.4% two or more races, and 0.14% some other. Choose one Hawaiian resident at random. What is the probability that he/she is a Native Hawaiian or Pacific Islander? Asian? White?

31. Crimes Committed The numbers show the number of crimes committed in a large city. If a crime is selected at random, find the probability that it is a motor vehicle theft. What is the probability that it is not an assault?

Theft	1375
Burglary of home or office	500
Motor vehicle theft	275
Assault	200
Robbery	125
Rape or homicide	25

Source: Based on FBI statistics.

32. Living Arrangements for Children Here are the living arrangements of children under 18 years old living in the United States in a recent year. Numbers are in thousands.

Both parents	51,823
Mother only	17,283
Father only	2,572
Neither parent	3,041

Choose one child at random; what is the probability that the child lives with both parents? With the mother present?

Source: Time Almanac.

33. Motor Vehicle Accidents During a recent year, there were 13.5 million automobile accidents, 5.2 million truck accidents, and 178,000 motorcycle accidents. If one accident is selected at random, find the probability that it is either a truck or motorcycle accident. What is the probability that it is not a truck accident?

Source: Based on data from the National Safety Council.

34. Federal Government Revenue The source of federal government revenue for a specific year is

50% from individual income taxes
32% from social insurance payroll taxes
10% from corporate income taxes
3% from excise taxes
5% other

If a revenue source is selected at random, what is the probability that it comes from individual or corporate income taxes?

Source: New York Times Almanac.

35. Selecting a Bill A box contains a $1 bill, a $5 bill, a $10 bill, and a $20 bill. A bill is selected at random, and it is not replaced; then a second bill is selected at random. Draw a tree diagram and determine the sample space.

36. Tossing Coins Draw a tree diagram and determine the sample space for tossing four coins.

37. Selecting Numbered Balls Four balls numbered 1 through 4 are placed in a box. A ball is selected at random, and its number is noted; then it is replaced. A second ball is selected at random, and its number is noted. Draw a tree diagram and determine the sample space.

38. Family Dinner Combinations A family special at a neighborhood restaurant offers dinner for four for $39.99. There are 3 appetizers available, 4 entrees, and 3 desserts from which to choose. The special includes one of each. Represent the possible dinner combinations with a tree diagram.

39. Required First-Year College Courses First-year students at a particular college must take one English class, one class in mathematics, a first-year seminar, and an elective. There are 2 English classes to choose from, 3 mathematics classes, 5 electives, and everyone takes the same first-year seminar. Represent the possible schedules, using a tree diagram.

40. Tossing a Coin and Rolling a Die A coin is tossed; if it falls heads up, it is tossed again. If it falls tails up, a die is rolled. Draw a tree diagram and determine the outcomes.

Extending the Concepts

41. Distribution of CEO Ages The distribution of ages of CEOs is as follows:

Age	Frequency
21–30	1
31–40	8
41–50	27
51–60	29
61–70	24
71–up	11

Source: Information based on *USA TODAY Snapshot.*

If a CEO is selected at random, find the probability that his or her age is

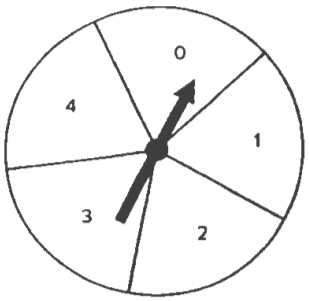
a. Between 31 and 40
b. Under 31
c. Over 30 and under 51
d. Under 31 or over 60

42. Tossing a Coin A person flipped a coin 100 times and obtained 73 heads. Can the person conclude that the coin was unbalanced?

43. Medical Treatment A medical doctor stated that with a certain treatment, a patient has a 50% chance of

recovering without surgery. That is, "Either he will get well or he won't get well." Comment on this statement.

44. Wheel Spinner The wheel spinner shown here is spun twice. Find the sample space, and then determine the probability of the following events.



a. An odd number on the first spin and an even number on the second spin (*Note:* 0 is considered even.)
b. A sum greater than 4
c. Even numbers on both spins
d. A sum that is odd
e. The same number on both spins

45. Tossing Coins Toss three coins 128 times and record the number of heads (0, 1, 2, or 3); then record your results with the theoretical probabilities. Compute the empirical probabilities of each.

46. Tossing Coins Toss two coins 100 times and record the number of heads (0, 1, 2). Compute the probabilities of each outcome, and compare these probabilities with the theoretical results.

47. Odds Odds are used in gambling games to make them fair. For example, if you rolled a die and won every time you rolled a 6, then you would win on average once every 6 times. So that the game is fair, the odds of 5 to 1 are given. This means that if you bet \$1 and won, you could win \$5. On average, you would win \$5 once in 6 rolls and lose \$1 on the other 5 rolls—hence the term *fair game*.

In most gambling games, the odds given are not fair. For example, if the odds of winning are really 20 to 1, the house might offer 15 to 1 in order to make a profit.

Odds can be expressed as a fraction or as a ratio, such as $\frac{5}{1}$, 5:1, or 5 to 1. Odds are computed in favor of the event or against the event. The formulas for odds are

$$\text{Odds in favor} = \frac{P(E)}{1 - P(E)}$$

$$\text{Odds against} = \frac{P(\overline{E})}{1 - P(\overline{E})}$$

In the die example,

$$\text{Odds in favor of a 6} = \frac{\frac{1}{6}}{\frac{5}{6}} = \frac{1}{5} \text{ or } 1:5$$

$$\text{Odds against a 6} = \frac{\frac{5}{6}}{\frac{1}{6}} = \frac{5}{1} \text{ or } 5:1$$

Find the odds in favor of and against each event.

a. Rolling a die and getting a 2
b. Rolling a die and getting an even number
c. Drawing a card from a deck and getting a spade
d. Drawing a card and getting a red card
e. Drawing a card and getting a queen
f. Tossing two coins and getting two tails
g. Tossing two coins and getting exactly one tail

4–2 The Addition Rules for Probability

OBJECTIVE 2

Find the probability of compound events, using the addition rules.

Many problems involve finding the probability of two or more events. For example, at a large political gathering, you might wish to know, for a person selected at random, the probability that the person is a female or is a Republican. In this case, there are three possibilities to consider:

1. The person is a female.
2. The person is a Republican.
3. The person is both a female and a Republican.

Consider another example. At the same gathering there are Republicans, Democrats, and Independents. If a person is selected at random, what is the probability that the person is a Democrat or an Independent? In this case, there are only two possibilities:

1. The person is a Democrat.
2. The person is an Independent.

The difference between the two examples is that in the first case, the person selected can be a female and a Republican at the same time. In the second case, the person selected cannot be both a Democrat and an Independent at the same time. In the second case,

Historical Note

The first book on probability, *The Book of Chance and Games*, was written by Jerome Cardan (1501–1576). Cardan was an astrologer, philosopher, physician, mathematician, and gambler. This book contained techniques on how to cheat and how to catch others at cheating.

the two events are said to be *mutually exclusive;* in the first case, they are not mutually exclusive.

Two events are mutually exclusive events or **disjoint events** if they cannot occur at the same time (i.e., they have no outcomes in common).

In another situation, the events of getting a 4 and getting a 6 when a single card is drawn from a deck are mutually exclusive events, since a single card cannot be both a 4 and a 6. On the other hand, the events of getting a 4 and getting a heart on a single draw are not mutually exclusive, since you can select the 4 of hearts when drawing a single card from an ordinary deck.

EXAMPLE 4–15 Determining Mutually Exclusive Events

Determine whether the two events are mutually exclusive. Explain your answer.

a. Randomly selecting a female student
Randomly selecting a student who is a junior

b. Randomly selecting a person with type A blood
Randomly selecting a person with type O blood

c. Rolling a die and getting an odd number
Rolling a die and getting a number less than 3

d. Randomly selecting a person who is under 21 years of age
Randomly selecting a person who is over 30 years of age

SOLUTION

a. These events are not mutually exclusive since a student can be both female and a junior.

b. These events are mutually exclusive since a person cannot have type A blood and type O blood at the same time.

c. These events are not mutually exclusive since the number 1 is both an odd number and a number less than 3.

d. These events are mutually exclusive since a person cannot be both under 21 and over 30 years of age at the same time.

EXAMPLE 4–16 Drawing a Card

Determine which events are mutually exclusive and which events are not mutually exclusive when a single card is drawn at random from a deck.

a. Getting a face card; getting a 6

b. Getting a face card; getting a heart

c. Getting a 7; getting a king

d. Getting a queen; getting a spade

SOLUTION

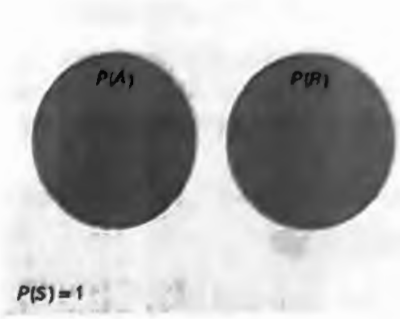
a. These events are mutually exclusive since you cannot draw one card that is both a face card (jack, queen, or king) and a card that is a 6 at the same time.

b. These events are not mutually exclusive since you can get one card that is a face card and is a heart: that is, a jack of hearts, a queen of hearts, or a king of hearts.

c. These events are mutually exclusive since you cannot get a single card that is both a 7 and a king.

d. These events are not mutually exclusive since you can get the queen of spades.

FIGURE 4–6
Venn Diagram for Addition Rule 1 When the Events Are Mutually Exclusive



Mutually exclusive events
$P(A \text{ or } B) = P(A) + P(B)$

The probability of two or more events can be determined by the *addition rules*. The first addition rule is used when the events are mutually exclusive.

Addition Rule 1

When two events A and B are mutually exclusive, the probability that A or B will occur is

$$P(A \text{ or } B) = P(A) + P(B)$$

Figure 4–6 shows a Venn diagram that represents two mutually exclusive events A and B. In this case, $P(A \text{ or } B) = P(A) + P(B)$, since these events are mutually exclusive and do not overlap. In other words, the probability of occurrence of event A or event B is the sum of the areas of the two circles.

EXAMPLE 4–17 Endangered Species

In the United States there are 59 different species of mammals that are endangered, 75 different species of birds that are endangered, and 68 species of fish that are endangered. If one animal is selected at random, find the probability that it is either a mammal or a fish.

Source: Based on information from the U.S. Fish and Wildlife Service.

SOLUTION

Since there are 59 species of mammals and 68 species of fish that are endangered and a total of 202 endangered species, $P(\text{mammal or fish}) = P(\text{mammal}) + P(\text{fish}) = \frac{59}{202} + \frac{68}{202} = \frac{127}{202} = 0.629$. The events are mutually exclusive.

EXAMPLE 4–18 Research and Development Employees

The corporate research and development centers for three local companies have the following numbers of employees:

U.S. Steel	110
Alcoa	750
Bayer Material Science	250

If a research employee is selected at random, find the probability that the employee is employed by U.S. Steel or Alcoa.

Source: Pittsburgh Tribune Review.

SOLUTION

$$P(\text{U.S. Steel or Alcoa}) = P(\text{U.S. Steel}) + P(\text{Alcoa})$$
$$= \frac{110}{1110} + \frac{750}{1110} = \frac{860}{1110} = \frac{86}{111} = 0.775$$

EXAMPLE 4–19 Favorite Ice Cream

In a survey, 8% of the respondents said that their favorite ice cream flavor is cookies and cream, and 6% like mint chocolate chip. If a person is selected at random, find the probability that her or his favorite ice cream flavor is either cookies and cream or mint chocolate chip.

Source: Rasmussen Report

SOLUTION

$$\begin{aligned} &P(\text{cookies and cream or mint chocolate chip}) \\ &\quad = P(\text{cookies and cream}) + P(\text{mint chocolate chip}) \\ &\quad = 0.08 + 0.06 = 0.14 = 14\% \end{aligned}$$

These events are mutually exclusive.

Historical Note

Venn diagrams were developed by mathematician John Venn (1834–1923) and are used in set theory and symbolic logic. They have been adapted to probability theory also. In set theory, the symbol $\cup$ represents the *union* of two sets, and $A \cup B$ corresponds to A or B. The symbol $\cap$ represents the *intersection* of two sets, and $A \cap B$ corresponds to A and B. Venn diagrams show only a general picture of the probability rules and do not portray all situations, such as $P(A) = 0$, accurately.

The probability rules can be extended to three or more events. For three mutually exclusive events A, B, and C,

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C)$$

When events are not mutually exclusive, addition rule 2 can be used to find the probability of the events.

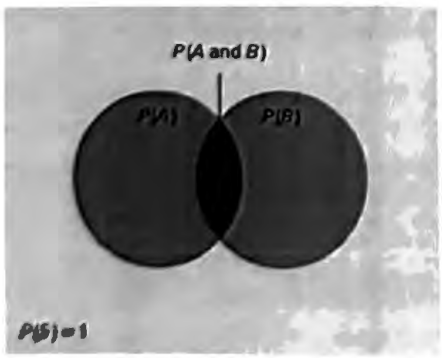
Addition Rule 2

If A and B are *not* mutually exclusive, then

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Note: This rule can also be used when the events are mutually exclusive, since $P(A \text{ and } B)$ will always equal 0. However, it is important to make a distinction between the two situations.

Figure 4–7 represents the probability of two events that are *not* mutually exclusive. In this case, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$. The area in the intersection or overlapping part of both circles corresponds to $P(A \text{ and } B)$; and when the area of circle A is added to the area of circle B, the overlapping part is counted twice. It must therefore be subtracted once to get the correct area or probability.



Nonmutually exclusive events
$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

FIGURE 4–7
Venn Diagram for Addition Rule 2 When Events Are Not Mutually Exclusive

EXAMPLE 4–20 Drawing a Card

A single card is drawn at random from an ordinary deck of cards. Find the probability that it is either a 9 or a diamond.

SOLUTION

There are 4 nines and 13 diamonds in a deck of cards, and one of the 9s is a diamond, so it is counted twice. Hence,

$$P(9 \text{ or } \blacklozenge) = P(9) + P(\blacklozenge) - P(9\blacklozenge) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13} \approx 0.308$$

EXAMPLE 4–21 Selecting a Medical Staff Person

In a hospital unit there are 8 nurses and 5 physicians; 7 nurses and 3 physicians are females. If a staff person is selected, find the probability that the subject is a nurse or a male.

SOLUTION

The sample space is shown here.

Staff	Females	Males	Total
Nurses	7	1	8
Physicians	3	2	5
Total	10	3	13

The probability is

$$P(\text{nurse or male}) = P(\text{nurse}) + P(\text{male}) - P(\text{male nurse})$$
$$= \frac{8}{13} + \frac{3}{13} - \frac{1}{13} = \frac{10}{13} \approx 0.769$$

EXAMPLE 4–22 Driving While Intoxicated

On New Year's Eve, the probability of a person driving while intoxicated is 0.32, the probability of a person having a driving accident is 0.09, and the probability of a person having a driving accident while intoxicated is 0.06. What is the probability of a person driving while intoxicated or having a driving accident?

SOLUTION

$$P(\text{intoxicated or accident}) = P(\text{intoxicated}) + P(\text{accident}) - P(\text{intoxicated and accident})$$
$$= 0.32 + 0.09 - 0.06 = 0.35$$

For three events that are *not* mutually exclusive,

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) - P(A \text{ and } B) - P(A \text{ and } C) - P(B \text{ and } C) + P(A \text{ and } B \text{ and } C)$$

See Exercises 23 and 24 in this section.

In summary, then, when the two events are mutually exclusive, use addition rule 1. When the events are not mutually exclusive, use addition rule 2.

Interesting Fact

Card Shuffling

How many times does a deck of cards need to be shuffled so that the cards are in random order? Actually, this question is not easy to answer since there are many variables. First, several different methods are used to shuffle a deck of cards. Some of the methods are the riffle method, the overhand method, the Corgi method, and the Faro method.

Another factor that needs to be considered is what is meant by the cards being in a random order. There are several statistical tests that can be used to determine if a deck of cards is randomized after several shuffles, but these tests give somewhat different results.

Two mathematicians, Persi Diaconis and Dave Bayer, concluded that a deck of cards starts to become random after 5 good shuffles and is completely random after 7 shuffles. However, a later study done by Trefthen concluded that only 6 shuffles are necessary. The difference was based on what is considered a randomized deck of cards.

Applying the Concepts 4–2

Which Pain Reliever Is Best?

Assume that following an injury you received from playing your favorite sport, you obtain and read information on new pain medications. In that information you read of a study that was conducted to test the side effects of two new pain medications. Use the following table to answer the questions and decide which, if any, of the two new pain medications you will use.

	Number of side effects in 12-week clinical trial		
Side effect	Placebo $n = 192$	Drug A $n = 186$	Drug B $n = 188$
Upper respiratory congestion	10	32	19
Sinus headache	11	25	32
Stomachache	2	46	12
Neurological headache	34	55	72
Cough	22	18	31
Lower respiratory congestion	2	5	1

1. How many subjects were in the study?
2. How long was the study?
3. What were the variables under study?
4. What type of variables are they, and what level of measurement are they on?
5. Are the numbers in the table exact figures?
6. What is the probability that a randomly selected person was receiving a placebo?
7. What is the probability that a person was receiving a placebo or drug A? Are these mutually exclusive events? What is the complement to this event?
8. What is the probability that a randomly selected person was receiving a placebo or experienced a neurological headache?
9. What is the probability that a randomly selected person was not receiving a placebo or experienced a sinus headache?

See page 254 for the answers.

Exercises 4–2

1. Define mutually exclusive events, and give an example of two events that are mutually exclusive and two events that are not mutually exclusive.

2. Explain briefly why addition rule 2 can be used when two events are mutually exclusive.

3. **Cards, Dice, and Students** Determine whether these events are mutually exclusive:
   a. Draw a card: get a spade and get a 6
   b. Roll a die: get a prime number (2, 3, 5)
   c. Roll two dice: get a sum of 7 or get a sum that is an even number
   d. Select a student at random in your class: get a male or get a sophomore

4. Determine whether these events are mutually exclusive.
   a. Roll two dice: Get a sum of 7 or get doubles.
   b. Select a student in your college: The student is a sophomore and the student is a business major.
   c. Select any course: It is a calculus course and it is an English course.
   d. Select a registered voter: The voter is a Republican and the voter is a Democrat.

5. **College Degrees Awarded** The table below represents the college degrees awarded in a recent academic year by gender.

	Bachelor's	Master's	Doctorate
Men	573,079	211,381	24,341
Women	775,424	301,264	21,683

Choose a degree at random. Find the probability that it is

a. A bachelor's degree

b. A doctorate or a degree awarded to a woman

c. A doctorate awarded to a woman

d. Not a master's degree

Source: www.nces.ed.gov

6. **Riding to School** The probability that John will drive to school is 0.37, the probability that he will ride with friends is 0.23, and the probability that his parents will take him is 0.4. He is not allowed to have passengers in the car when he is driving. What is the probability that John will have company on the way to school?

7. **Medical Specialities** The following doctors were observed on staff at a local hospital.

	MD	Doctor of Osteopathy
Pathology	6	1
Pediatrics	7	2
Orthopedics	20	2

Choose one doctor at random; what is the probability that

a. She is a pathologist?

b. He is an orthopedist or an MD?

8. **U.S. Population** The data show the U.S. population by age.

Under 20 years	27.0%
20 years and over	73.0
65 years and over	13.1

Choose one person from the United States at random. Find the probability that the person is

a. From 20 years to 64 years

b. Under 20 or 65 and over

c. Not 65 and over

9. **Snack Foods** In a meeting room in a dormitory there are 8 bags of potato chips, 5 bags of popcorn, 2 bags of pretzels, and 1 bag of cheese puffs. If a student selects 1 bag at random, find the probability that it is a bag of potato chips or a bag of pretzels.

10. **Selecting a Movie** A media rental store rented the following number of movie titles in each of these categories: 170 horror, 230 drama, 120 mystery, 310 romance, and 150 comedy. If a person selects a movie to rent, find the probability that it is a romance or a comedy. Is this event likely or unlikely to occur? Explain your answer.

11. **Pizza Sales** A pizza restaurant sold 24 cheese pizzas and 16 pizzas with one or more toppings. Twelve of the cheese pizzas were eaten at work, and 10 of the pizzas with one or more toppings were eaten at work. If a pizza was selected at random, find the probability of each:

a. It was a cheese pizza eaten at work.

b. It was a pizza with either one or more toppings, and it was not eaten at work.

c. It was a cheese pizza, or it was a pizza eaten at work.

12. **Selecting a Book** At a used-book sale, 100 books are adult books and 160 are children's books. Of the adult books, 70 are nonfiction while 60 of the children's books are nonfiction. If a book is selected at random, find the probability that it is

a. Fiction

b. Not a children's nonfiction book

c. An adult book or a children's nonfiction book

13. **Young Adult Residences** According to the Bureau of the Census, the following statistics describe the number (in thousands) of young adults living at home or in a dormitory in the year 2004.

	Ages 18–24	Ages 25–34
Male	7922	2534
Female	5779	995

Source: World Almanac.

Choose one student at random. Find the probability that the student is

a. A female student aged 25–34 years

b. Male or aged 18–24 years

c. Under 25 years of age and not male

14. **Endangered Species** The chart below shows the numbers of endangered and threatened species both here in the United States and abroad.

	Endangered		Threatened	
	United States	Foreign	United States	Foreign
Mammals	68	251	10	20
Birds	77	175	13	6
Reptiles	14	64	22	16
Amphibians	11	8	10	1

Source: www.infoplease.com

Choose one species at random. Find the probability that it is

a. Threatened and in the United States

b. An endangered foreign bird

c. A mammal or a threatened foreign species

15. **Multiple Births** The number of multiple births in the United States for a recent year indicated that there were 128,665 sets of twins, 7110 sets of triplets, 468 sets of quadruplets, and 85 sets of quintuplets. Choose one set of siblings at random.

a. Find the probability that it represented more than two babies.

b. Find the probability that it represented quads or quints.

c. Now choose one baby from these multiple births. What is the probability that the baby was a triplet?

16. Licensed Drivers in the United States In a recent year there were the following numbers (in thousands) of licensed drivers in the United States.

	Male	Female
Age 19 and under	4746	4517
Age 20	1625	1553
Age 21	1679	1627

Source: World Almanac

Choose one driver at random. Find the probability that the driver is

a. Male and 19 years or under

b. Age 20 or female

c. At least 20 years old

17. Prison Education In a federal prison, inmates can select to complete high school, take college courses, or do neither. The following survey results were obtained using ages of the inmates.

Age	High School Courses	College Courses	Neither
Under 30	53	107	450
30 and over	27	32	367

If a prisoner is selected at random, find these probabilities:

a. The prisoner does not take classes.

b. The prisoner is under 30 and is taking either a high school class or a college class.

c. The prisoner is over 30 and is taking either a high school class or a college class.

18. Mail Delivery A local postal carrier distributes first-class letters, advertisements, and magazines. For a certain day, she distributed the following numbers of each type of item.

Delivered to	First-class letters	Ads	Magazines
Home	325	406	203
Business	732	1021	97

If an item of mail is selected at random, find these probabilities.

a. The item went to a home.

b. The item was an ad, or it went to a business.

c. The item was a first-class letter, or it went to a home.

19. Medical Tests on Emergency Patients The frequency distribution shown here illustrates the number of medical tests conducted on 30 randomly selected emergency patients.

Number of tests performed	Number of patients
0	12
1	8
2	2
3	3
4 or more	5

If a patient is selected at random, find these probabilities.

a. The patient has had exactly 2 tests done.

b. The patient has had at least 2 tests done.

c. The patient has had at most 3 tests done.

d. The patient has had 3 or fewer tests done.

e. The patient has had 1 or 2 tests done.

20. College Fundraiser A social organization of 32 members sold college sweatshirts as a fundraiser. The results of their sale are shown below.

No. of sweatshirts	No. of students
0	2
1–5	13
6–10	8
11–15	4
16–20	4
20+	1

Choose one student at random. Find the probability that the student sold

a. More than 10 sweatshirts

b. At least one sweatshirt

c. 1–5 or more than 15 sweatshirts

21. Emergency Room Tests The frequency distribution shows the number of medical tests conducted on 30 randomly selected emergency room patients.

Number of tests performed	Number of patients
0	11
1	9
2	5
3	4
4 or more	1

If a patient is selected at random, find these probabilities:

a. The patient had exactly 3 tests done.

b. The patient had at most 2 tests done.

c. The patient has 1 or 2 tests done.

d. The patient had fewer than 3 tests done.

e. The patient had at least 3 tests done.

22. Medical Patients A recent study of 300 patients found that of 100 alcoholic patients, 87 had elevated cholesterol levels, and of 200 nonalcoholic patients, 43 had elevated cholesterol levels. If a patient is selected at random, find the probability that the patient is the following.

a. An alcoholic with elevated cholesterol level

b. A nonalcoholic

c. A nonalcoholic with nonelevated cholesterol level

23. Selecting a Card If one card is drawn from an ordinary deck of cards, find the probability of getting each event:

a. A 7 or an 8 or a 9

b. A spade or a queen or a king

c. A club or a face card

d. An ace or a diamond or a heart

e. A 9 or a 10 or a spade or a club

24. Rolling Die Two dice are rolled. Find the probability of getting

a. A sum of 8, 9, or 10

b. Doubles or a sum of 7

c. A sum greater than 9 or less than 4

d. Based on the answers to *a*, *b*, and *c*, which is least likely to occur?

25. Apple Production For a recent year, about 11 billion pounds of apples were harvested. About 4.4 billion pounds of apples were made into apple juice, about 1 billion pounds of apples were made into apple sauce, and 1 billion pounds of apples were used for other commercial purposes. If 1 billion pounds of apples were selected at random, what is the probability that the apples were used for apple juice or applesauce?

Source: International Apple Institute

26. Rolling Dice Three dice are rolled. Find the probability of getting

a. Triples *b.* A sum of 5

Extending the Concepts

27. Purchasing a Pizza The probability that a customer selects a pizza with mushrooms or pepperoni is 0.55, and the probability that the customer selects only mushrooms is 0.32. If the probability that he or she selects only pepperoni is 0.17, find the probability of the customer selecting both items.

28. Building a New Home In building new homes, a contractor finds that the probability of a home buyer selecting a two-car garage is 0.70 and of selecting a one-car garage is 0.20. Find the probability that the buyer will select no garage. The builder does not build houses with three-car or more garages.

29. In Exercise 28, find the probability that the buyer will not want a two-car garage.

30. Suppose that $P(A) = 0.42$, $P(B) = 0.38$, and $P(A \text{ or } B) = 0.70$. Are A and B mutually exclusive? Explain.

31. The probability of event A occurring is $m/(2m + n)$, and the probability of event B occurring is $n/(2m + n)$. Find the probability of A or B occurring if the events are mutually exclusive.

32. Events A and B are mutually exclusive with $P(A)$ equal to 0.392 and $P(A \text{ or } B)$ equal to 0.653. Find

a. $P(B)$

b. $P(\text{not } A)$

c. $P(A \text{ and } B)$

Technology Step by Step

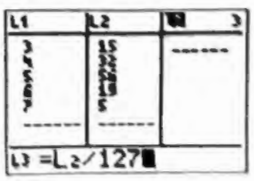
TI-84 Plus Step by Step

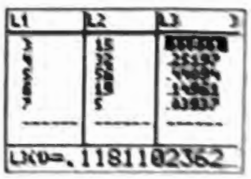
To construct a relative frequency table:

1. Enter the data values in L_1 and the frequencies in L_2.
2. Move the cursor to the top of the L_3 column so that L_3 is highlighted.
3. Type L_2 divided by the sample size, then press **ENTER**.

Example TI4–1

Construct a relative frequency table for the knee replacement data from Example 4–14.





4–3 The Multiplication Rules and Conditional Probability

Section 4–2 showed that the addition rules are used to compute probabilities for mutually exclusive and non-mutually exclusive events. This section introduces the multiplication rules.

The Multiplication Rules

OBJECTIVE 3

Find the probability of compound events, using the multiplication rules

The *multiplication rules* can be used to find the probability of two or more events that occur in sequence. For example, if you toss a coin and then roll a die, you can find the probability of getting a head on the coin *and* a 4 on the die. These two events are said to be *independent* since the outcome of the first event (tossing a coin) does not affect the probability outcome of the second event (rolling a die).

Two events *A* and *B* are Independent events if the fact that *A* occurs does not affect the probability of *B* occurring.

Here are other examples of independent events:

Rolling a die and getting a 6, and then rolling a second die and getting a 3.

Drawing a card from a deck and getting a queen, replacing it, and drawing a second card and getting a queen.

To find the probability of two independent events that occur in sequence, you must find the probability of each event occurring separately and then multiply the answers. For example, if a coin is tossed twice, the probability of getting two heads is $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. This result can be verified by looking at the sample space HH, HT, TH, TT. Then $P(\text{HH}) = \frac{1}{4}$.

Multiplication Rule 1

When two events are independent, the probability of both occurring is

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

EXAMPLE 4–23 Tossing a Coin

A coin is flipped and a die is rolled. Find the probability of getting a head on the coin and a 4 on the die.

SOLUTION

The sample space for the coin is H, T; and for the die it is 1, 2, 3, 4, 5, 6.

$$P(\text{head and } 4) = P(\text{head}) \cdot P(4) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12} \approx 0.083$$

The problem in Example 4–23 can also be solved by using the sample space

H1 H2 H3 H4 H5 H6 T1 T2 T3 T4 T5 T6

The solution is $\frac{1}{12}$, since there is only one way to get the head-4 outcome.

EXAMPLE 4–24 Drawing a Card

A card is drawn from a deck and replaced; then a second card is drawn. Find the probability of getting a king and then a 7.

SOLUTION

The probability of getting a king is $\frac{4}{52}$, and the probability of getting a 7 is $\frac{4}{52}$; hence, the probability of getting a king and then a 7 is

$$P(\text{king and then a 7}) = \frac{4}{52} \cdot \frac{4}{52} = \frac{16}{2704} = \frac{1}{169} \approx 0.006.$$

EXAMPLE 4–25 Selecting a Colored Ball

An urn contains 2 red balls, 5 blue balls, and 3 white balls. A ball is selected and its color is noted. Then it is replaced. A second ball is selected and its color is noted. Find the probability of each of these events.

a. Selecting 3 blue balls

b. Selecting 1 white ball and then a red ball

c. Selecting 2 blue balls and then one white ball

SOLUTION

a. $P(\text{blue and blue and blue}) = P(\text{blue}) \cdot P(\text{blue}) \cdot P(\text{blue}) = \frac{5}{10} \cdot \frac{5}{10} \cdot \frac{5}{10} = \frac{125}{1000} = \frac{1}{8} = 0.125$

b. $P(\text{white and red}) = P(\text{white}) \cdot P(\text{red}) = \frac{3}{10} \cdot \frac{2}{10} = \frac{6}{100} = \frac{3}{50} = 0.06$

c. $P(\text{blue and blue and white}) = P(\text{blue}) \cdot P(\text{blue}) \cdot P(\text{white}) = \frac{5}{10} \cdot \frac{5}{10} \cdot \frac{3}{10} = \frac{75}{1000} = \frac{3}{40} = 0.075$

Multiplication rule 1 can be extended to three or more independent events by using the formula

$$P(A \text{ and } B \text{ and } C \text{ and } \ldots \text{ and } K) = P(A) \cdot P(B) \cdot P(C) \cdots P(K)$$

When a small sample is selected from a large population and the subjects are not replaced, the probability of the event occurring changes so slightly that for the most part, it is considered to remain the same. Examples 4–26 and 4–27 illustrate this concept.

EXAMPLE 4–26 Bank Robberies

It was found that 3 out of every 4 people who commit a bank robbery are apprehended. (*Christian Science Monitor*). If 3 bank robberies are selected at random, find the probability that all three robbers will be apprehended.

SOLUTION

Let R be the probability that a bank robber is apprehended. Then

$$P(R \text{ and } R \text{ and } R) = P(R) \cdot P(R) \cdot P(R) = \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} = \frac{27}{64} \approx 0.422$$

There is about a 42% chance that all 3 robbers will be apprehended.

EXAMPLE 4–27 Male Color Blindness

Approximately 9% of men have a type of color blindness that prevents them from distinguishing between red and green. If 3 men are selected at random, find the probability that all of them will have this type of red-green color blindness.

Source: USA TODAY.

SOLUTION

Let C denote red-green color blindness. Then

$$\begin{aligned} P(C \text{ and } C \text{ and } C) &= P(C) \cdot P(C) \cdot P(C) \\ &= (0.09)(0.09)(0.09) \\ &= 0.000729 \end{aligned}$$

Hence, the rounded probability is 0.0007.

There is a 0.07% chance that all three men selected will have this type of red-green color blindness.

In Examples 4–23 through 4–27, the events were independent of one another, since the occurrence of the first event in no way affected the outcome of the second event. On the other hand, when the occurrence of the first event changes the probability of the occurrence of the second event, the two events are said to be *dependent*. For example, suppose a card is drawn from a deck and *not* replaced, and then a second card is drawn. What is the probability of selecting an ace on the first card and a king on the second card?

Before an answer to the question can be given, you must realize that the events are dependent. The probability of selecting an ace on the first draw is $\frac{4}{52}$. If that card is *not* replaced, the probability of selecting a king on the second card is $\frac{4}{51}$, since there are 4 kings and 51 cards remaining. The outcome of the first draw has affected the outcome of the second draw.

Dependent events are formally defined now.

When the outcome or occurrence of the first event affects the outcome or occurrence of the second event in such a way that the probability is changed, the events are said to be **dependent events.**

Here are some examples of dependent events:

Drawing a card from a deck, not replacing it, and then drawing a second card

Selecting a ball from an urn, not replacing it, and then selecting a second ball

Being a lifeguard and getting a suntan

Having high grades and getting a scholarship

Parking in a no-parking zone and getting a parking ticket

To find probabilities when events are dependent, use the multiplication rule with a modification in notation. For the problem just discussed, the probability of getting an ace on the first draw is $\frac{4}{52}$, and the probability of getting a king on the second draw is $\frac{4}{51}$. By the multiplication rule, the probability of both events occurring is

$$\frac{4}{52} \cdot \frac{4}{51} = \frac{16}{2652} = \frac{4}{663} \approx 0.006$$

The event of getting a king on the second draw *given* that an ace was drawn the first time is called a *conditional probability*.

The **conditional probability** of an event B in relationship to an event A is the probability that event B occurs after event A has already occurred. The notation for conditional probability is $P(B|A)$. This notation does not mean that B is divided by A; rather, it means the probability that event B occurs given that event A has already occurred. In the card example, $P(B|A)$ is the probability that the second card is a king given that the first card is an ace, and it is equal to $\frac{4}{51}$ since the first card was *not* replaced.

Multiplication Rule 2

When two events are dependent, the probability of both occurring is

$$P(A \text{ and } B) = P(A) \cdot P(B|A)$$

EXAMPLE 4–28 Unemployed Workers

For a specific year, 5.2% of U.S. workers were unemployed. During that time, 33% of those who were unemployed received unemployment benefits. If a person is selected at random, find the probability that she or he received unemployment benefits if the person is unemployed.

Source: Bureau of Labor Statistics

SOLUTION

$P(\text{unemployed benefits and unemployed}) = P(U) \cdot P(B|U) = (0.052)(0.33) = 0.017$

There is a 0.017 probability that a person is unemployed and receiving unemployment benefits.

EXAMPLE 4–29 Homeowner's and Automobile Insurance

World Wide Insurance Company found that 53% of the residents of a city had homeowner's insurance (H) with the company. Of these clients, 27% also had automobile insurance (A) with the company. If a resident is selected at random, find the probability that the resident has both homeowner's and automobile insurance with World Wide Insurance Company.

SOLUTION

$$P(\text{H and A}) = P(\text{H}) \cdot P(\text{A}|\text{H}) = (0.53)(0.27) = 0.1431 \approx 0.143$$

There is about a 14.3% probability that a resident has both homeowner's and automobile insurance with World Wide Insurance Company.

This multiplication rule can be extended to three or more events, as shown in Example 4–30.

EXAMPLE 4–30 Drawing Cards

Three cards are drawn from an ordinary deck and not replaced. Find the probability of these events.

a. Getting 3 jacks

b. Getting an ace, a king, and a queen in order

c. Getting a club, a spade, and a heart in order

d. Getting 3 clubs

SOLUTION

a. $P(3 \text{ jacks}) = \frac{4}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} = \frac{24}{132{,}600} = \frac{1}{5525} \approx 0.0002$

b. $P(\text{ace and king and queen}) = \frac{4}{52} \cdot \frac{4}{51} \cdot \frac{4}{50} = \frac{64}{132{,}600} = \frac{8}{16{,}575} \approx 0.0005$

c. $P(\text{club and spade and heart}) = \frac{13}{52} \cdot \frac{13}{51} \cdot \frac{13}{50} = \frac{2197}{132{,}600} = \frac{169}{10{,}200} \approx 0.017$

d. $P(3 \text{ clubs}) = \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} = \frac{1716}{132{,}600} = \frac{11}{850} \approx 0.013$

Tree diagrams can be used as an aid to finding the solution to probability problems when the events are sequential. Example 4–31 illustrates the use of tree diagrams.

EXAMPLE 4–31 Selecting Colored Balls

Box 1 contains 2 red balls and 1 blue ball. Box 2 contains 3 blue balls and 1 red ball. A coin is tossed. If it falls heads up, box 1 is selected and a ball is drawn. If it falls tails up, box 2 is selected and a ball is drawn. Find the probability of selecting a red ball.

SOLUTION

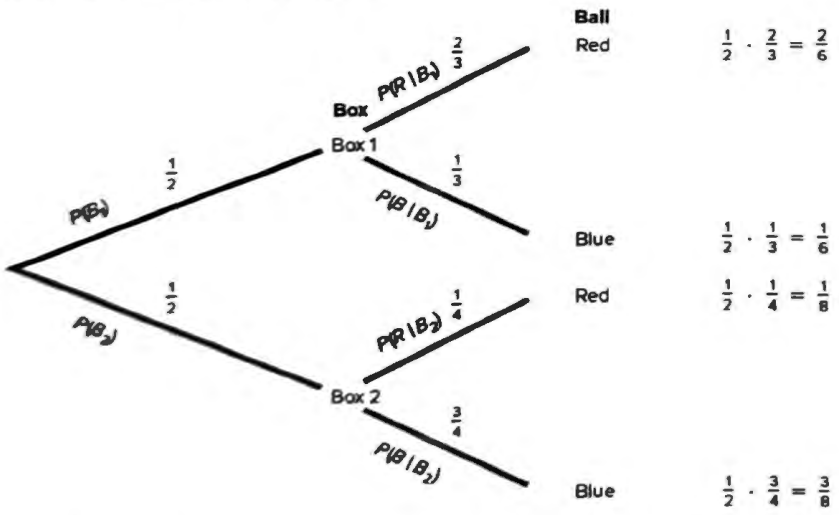
The first two branches designate the selection of either box 1 or box 2. Then from box 1, either a red ball or a blue ball can be selected. Likewise, a red ball or blue ball can be selected from box 2. Hence, a tree diagram of the example is shown in Figure 4–8.

Next determine the probabilities for each branch. Since a coin is being tossed for the box selection, each branch has a probability of $\frac{1}{2}$, that is, heads for box 1 or tails for box 2. The probabilities for the second branches are found by using the basic probability rule. For example, if box 1 is selected and there are 2 red balls and 1 blue ball, the probability of selecting a red ball is $\frac{2}{3}$ and the probability of selecting a blue ball is $\frac{1}{3}$. If box 2 is selected and it contains 3 blue balls and 1 red ball, then the probability of selecting a red ball is $\frac{1}{4}$ and the probability of selecting a blue ball is $\frac{3}{4}$.

Next multiply the probability for each outcome, using the rule $P(A \text{ and } B) = P(A) \cdot P(B|A)$. For example, the probability of selecting box 1 and selecting a red ball is $\frac{1}{2} \cdot \frac{2}{3} = \frac{2}{6}$. The probability of selecting box 1 and a blue ball is $\frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$. The probability of selecting box 2 and selecting a red ball is $\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$. The probability of selecting box 2 and a blue ball is $\frac{1}{2} \cdot \frac{3}{4} = \frac{3}{8}$. (Note that the sum of these probabilities is 1.)

Finally, a red ball can be selected from either box 1 or box 2 so $P(\text{red}) = \frac{2}{6} + \frac{1}{8} = \frac{8}{24} + \frac{3}{24} = \frac{11}{24}$.

FIGURE 4–8 Tree Diagram for Example 4–31



Tree diagrams can be used when the events are independent or dependent, and they can also be used for sequences of three or more events.

OBJECTIVE 4

Find the conditional probability of an event

Conditional Probability

The conditional probability of an event B in relationship to an event A was defined as the probability that event B occurs after event A has already occurred.

The conditional probability of an event can be found by dividing both sides of the equation for multiplication rule 2 by $P(A)$, as shown:

$$P(A \text{ and } B) = P(A) \cdot P(B|A)$$

$$\frac{P(A \text{ and } B)}{P(A)} = \frac{P(A) \cdot P(B|A)}{P(A)}$$

$$\frac{P(A \text{ and } B)}{P(A)} = P(B|A)$$

Formula for Conditional Probability

The probability that the second event B occurs given that the first event A has occurred can be found by dividing the probability that both events occurred by the probability that the first event has occurred. The formula is

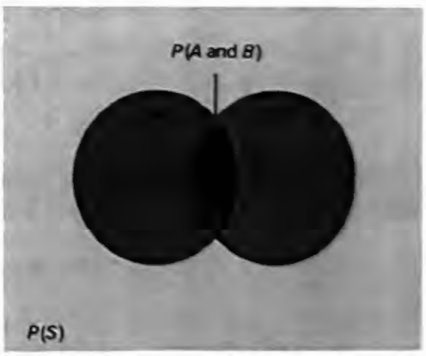
$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

The Venn diagram for conditional probability is shown in Figure 4–9. In this case,

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

which is represented by the area in the intersection or overlapping part of the circles A and B, divided by the area of circle A. The reasoning here is that if you assume A has occurred, then A becomes the sample space for the next calculation and is the denominator of the probability fraction $P(A \text{ and } B)/P(A)$. The numerator $P(A \text{ and } B)$ represents the probability of the part of B that is contained in A. Hence, $P(A \text{ and } B)$ becomes the numerator of the probability fraction $P(A \text{ and } B)/P(A)$. Imposing a condition reduces the sample space.

FIGURE 4–9
Venn Diagram for Conditional Probability



Examples 4–32, 4–33, and 4–34 illustrate the use of this rule.

EXAMPLE 4–32 Selecting Colored Chips

A box contains black chips and white chips. A person selects two chips without replacement. If the probability of selecting a black chip *and* a white chip is $\frac{15}{56}$ and the probability of selecting a black chip on the first draw is $\frac{5}{8}$, find the probability of selecting the white chip on the second draw, *given* that the first chip selected was a black chip.

SOLUTION

Let

B = selecting a black chip $\quad$ W = selecting a white chip

Then

$$P(W|B) = \frac{P(B \text{ and } W)}{P(B)} = \frac{15/56}{3/8}$$

$$= \frac{15}{56} \div \frac{3}{8} = \frac{15}{56} \cdot \frac{8}{3} = \frac{\overset{5}{\cancel{15}}}{\underset{7}{\cancel{56}}} \cdot \frac{\overset{1}{\cancel{8}}}{\underset{1}{\cancel{3}}} = \frac{5}{7} \approx 0.714$$

Hence, the probability of selecting a white chip on the second draw given that the first chip selected was black is $\frac{5}{7} \approx 0.714$.

EXAMPLE 4–33 Parking Tickets

The probability that Sam parks in a no-parking zone *and* gets a parking ticket is 0.06, and the probability that Sam cannot find a legal parking space and has to park in the no-parking zone is 0.20. On Tuesday, Sam arrives at school and has to park in a no-parking zone. Find the probability that he will get a parking ticket.

SOLUTION

Let

N = parking in a no-parking zone $\quad$ T = getting a ticket

Then

$$P(T|N) = \frac{P(N \text{ and } T)}{P(N)} = \frac{0.06}{0.20} = 0.30$$

Hence, Sam has a 0.30 probability or 30% chance of getting a parking ticket, given that he parked in a no-parking zone.

The conditional probability of events occurring can also be computed when the data are given in table form, as shown in Example 4–34.

EXAMPLE 4–34 Survey on Women in the Military

A recent survey asked 100 people if they thought women in the armed forces should be permitted to participate in combat. The results of the survey are shown.

Male	32	18	50
Female	8	42	50
Total	40	60	100

Find these probabilities.

a. The respondent answered yes, given that the respondent was a female.

b. The respondent was a male, given that the respondent answered no.

SOLUTION

Let

M = respondent was a male Y = respondent answered yes

F = respondent was a female N = respondent answered no

a. The problem is to find $P(Y|F)$. The rule states

$$P(Y|F) = \frac{P(F \text{ and } Y)}{P(F)}$$

The probability $P(F \text{ and } Y)$ is the number of females who responded yes, divided by the total number of respondents:

$$P(F \text{ and } Y) = \frac{8}{100}$$

The probability $P(F)$ is the probability of selecting a female:

$$P(F) = \frac{50}{100}$$

Then

$$P(Y|F) = \frac{P(F \text{ and } Y)}{P(F)} = \frac{8/100}{50/100}$$

$$= \frac{8}{100} \div \frac{50}{100} = \frac{\overset{4}{\cancel{8}}}{\underset{1}{\cancel{100}}} \cdot \frac{\overset{1}{\cancel{100}}}{\underset{25}{\cancel{50}}} = \frac{4}{25} = 0.16$$

b. The problem is to find $P(M|N)$.

$$P(M|N) = \frac{P(N \text{ and } M)}{P(N)} = \frac{18/100}{60/100}$$

$$= \frac{18}{100} \div \frac{60}{100} = \frac{\overset{3}{\cancel{18}}}{\underset{1}{\cancel{100}}} \cdot \frac{\overset{1}{\cancel{100}}}{\underset{10}{\cancel{60}}} = \frac{3}{10} = 0.3$$

Probabilities for "At Least"

The multiplication rules can be used with the complementary event rule (Section 4–1) to simplify solving probability problems involving "at least." Examples 4–35, 4–36, and 4–37 illustrate how this is done.

EXAMPLE 4–35 Drawing Cards

A person selects 3 cards from an ordinary deck and replaces each card after it is drawn. Find the probability that the person will get at least one heart.

SOLUTION

It is much easier to find the probability that the person will not select a heart in three draws and subtract this value from 1. To do the problem directly, you would have to find the probability of selecting 1 heart, 2 hearts, and 3 hearts and then add the results.

Let

E = at least 1 heart is drawn and $\overline{E}$ = no hearts are drawn

$$P(\overline{E}) = \frac{39}{52} \cdot \frac{39}{52} \cdot \frac{39}{52} = \frac{3}{4} \cdot \frac{3}{4} \cdot \frac{3}{4} = \frac{27}{64}$$

$$P(\overline{E}) = 1 - P(\overline{E})$$

$$= 1 - \frac{27}{64} = \frac{37}{64} \approx 0.578 = 57.8\%$$

Hence, a person will select at least one heart about 57.8% of the time.

EXAMPLE 4–36 Rolling a Die

A single die is rolled 4 times. Find the probability of getting at least one 6.

SOLUTION

It is easier to find the probability of the complement of the event, which is no 6s. Then subtract this probability from 1 in order to find the probability of getting at least one 6.

$$P(\text{at least one 6}) = 1 - P(\text{no 6s})$$

$$= 1 - \left(\frac{5}{6}\right)^4$$

$$= 1 - \frac{625}{1296}$$

$$= \frac{671}{1296} \approx 0.518$$

There is about a 51.8% chance of getting at least one 6 when a die is rolled four times.

EXAMPLE 4–37 Ties

The Neckware Association of America reported that 3% of ties sold in the United States are bow ties. If 4 customers who purchased a tie are randomly selected, find the probability that at least 1 purchased a bow tie.

SOLUTION

Let E = at least 1 bow tie is purchased and $\overline{E}$ = no bow ties are purchased. Then

$$P(E) = 0.03 \quad \text{and} \quad P(\overline{E}) = 1 - 0.03 = 0.97$$

$P(\text{no bow ties are purchased}) = (0.97)(0.97)(0.97)(0.97) \approx 0.885$; hence,
$P(\text{at least one bow tie is purchased}) = 1 - 0.885 = 0.115$.

There is an 11.5% chance of a person purchasing at least one bow tie.

Applying the Concepts 4–3

Guilty or Innocent?

In July 1964, an elderly woman was mugged in Costa Mesa, California. In the vicinity of the crime a tall, bearded man sat waiting in a yellow car. Shortly after the crime was committed, a young, tall woman, wearing her blond hair in a ponytail, was seen running from the scene of the crime and getting into the car, which sped off. The police broadcast a description of the suspected muggers. Soon afterward, a couple fitting the description was arrested and convicted of the crime. Although the evidence in the case was largely circumstantial, the two people arrested were nonetheless

convicted of the crime. The prosecutor based his entire case on basic probability theory, showing the unlikeness of another couple being in that area while having all the same characteristics that the elderly woman described. The following probabilities were used.

Characteristic	Assumed probability
Drives yellow car	1 out of 12
Man over 6 feet tall	1 out of 10
Man wearing tennis shoes	1 out of 4
Man with beard	1 out of 11
Woman with blond hair	1 out of 3
Woman with hair in a ponytail	1 out of 13
Woman over 6 feet tall	1 out of 100

1. Compute the probability of another couple being in that area with the same characteristics.
2. Would you use the addition or multiplication rule? Why?
3. Are the characteristics independent or dependent?
4. How are the computations affected by the assumption of independence or dependence?
5. Should any court case be based solely on probabilities?
6. Would you convict the couple who was arrested even if there were no eyewitnesses?
7. Comment on why in today's justice system no person can be convicted solely on the results of probabilities.
8. In actuality, aren't most court cases based on uncalculated probabilities?

See page 254 for the answers.

Exercises 4–3

1. State which events are independent and which are dependent.
   a. Tossing a coin and drawing a card from a deck
   b. Drawing a ball from an urn, not replacing it, and then drawing a second ball
   c. Getting a raise in salary and purchasing a new car
   d. Driving on ice and having an accident
2. State which events are independent and which are dependent.
   a. Having a large shoe size and having a high IQ
   b. A father being left-handed and a daughter being left-handed
   c. Smoking excessively and having lung cancer
   d. Eating an excessive amount of ice cream and smoking an excessive amount of cigarettes
3. **Video and Computer Games** Sixty-nine percent of U.S. heads of household play video or computer games. Choose 4 heads of household at random. Find the probability that
   a. None play video or computer games.
   b. All four do.

   Source: www.theesa.com
4. **Seat Belt Use** The Gallup Poll reported that 52% of Americans used a seat belt the last time they got into a car. If 4 people are selected at random, find the probability that they all used a seat belt the last time they got into a car.

   Source: 100% American.
5. **Automobile Sales** An automobile salesperson finds the probability of making a sale is 0.21. If she talks to 4 customers, find the probability that she will make 4 sales. Is the event likely or unlikely to occur? Explain your answer.
6. **Prison Populations** If 25% of U.S. federal prison inmates are not U.S. citizens, find the probability that 2 randomly selected federal prison inmates will not be U.S. citizens.

   Source: Harper's Index.
7. **Government Employees** In 2013 about 66% of full-time law enforcement workers were sworn officers, and of those, 88.4% were male. Females however make up 60.7% of civilian employees. Choose one law enforcement worker at random and find the following.
   a. The probability that she is a female sworn officer
   b. The probability that he is a male civilian employee
   c. The probability that he or she is male or a civilian employee

   Source: World Almanac.

8. **Working Women and Computer Use** It is reported that 72% of working women use computers at work. Choose 5 working women at random. Find
   a. The probability that at least 1 doesn't use a computer at work
   b. The probability that all 5 use a computer in their jobs

   Source: www.infoplease.com

9. **Female Prison Inmates** Seventy-five percent of female prison inmates are mothers. If 3 female prison inmates are selected at random, what is the probability that none are mothers?

   Source: Chicago Legal Aid to Incarcerated Mothers

10. **Selecting Marbles** A bag contains 9 red marbles, 8 white marbles, and 6 blue marbles. Randomly choose two marbles, one at a time, and without replacement. Find the following.
    a. The probability that the first marble is red and the second is white
    b. The probability that both are the same color
    c. The probability that the second marble is blue

11. **Smart TVs** Smart TVs have seen success in the United States market. During the 2nd quarter of 2015 45% of TVs sold in the United States were Smart TVs. That's an increase of 11% from 2014. Choose three households and find the probability that
    a. None of the 3 households had a Smart TV
    b. All 3 households had a Smart TV
    c. At least 1 of the 3 households had a Smart TV

    Source: The NPD Group Connected Intelligence Home Entertainment Report

12. **Flashlight Batteries** A flashlight has 6 batteries, 2 of which are defective. If 2 are selected at random without replacement, find the probability that both are defective.

13. **Drawing a Card** Four cards are drawn from a desk *without* replacement. Find these probabilities.
    a. All cards are jacks.
    b. All cards are black cards.
    c. All cards are hearts.

14. **Scientific Study** In a scientific study there are 8 guinea pigs, 5 of which are pregnant. If 3 are selected at random without replacement, find the probability that all are pregnant.

15. **Drawing Cards** If two cards are selected from a standard deck of 52 cards and are not replaced after each draw, find these probabilities.
    a. Both are 9s
    b. Both cards are the same suit.
    c. Both cards are spades.

16. **Winning a Door Prize** At a gathering consisting of 10 men and 20 women, two door prizes are awarded. Find the probability that both prizes are won by men. The winning ticket is not replaced. Would you consider this event likely or unlikely to occur?

17. **Defective Batteries** In a box of 12 batteries, 2 are dead. If 2 batteries are selected at random for a flashlight, find the probability that both are dead. Would you consider this event likely or unlikely?

18. **Sales** A manufacturer makes two models of an item: model I, which accounts for 80% of unit sales, and model II, which accounts for 20% of unit sales. Because of defects, the manufacturer has to replace (or exchange) 10% of its model I and 18% of its model II. If a model is selected at random, find the probability that it will be defective.

19. **Student Financial Aid** In a recent year 8,073,000 male students and 10,980,000 female students were enrolled as undergraduates. Receiving aid were 60.6% of the male students and 65.2% of the female students. Of those receiving aid, 44.8% of the males got federal aid and 50.4% of the females got federal aid. Choose 1 student at random. (*Hint:* Make a tree diagram.) Find the probability that the student is
    a. A male student without aid
    b. A male student, given that the student has aid
    c. A female student or a student who receives federal aid

    Source: www.nces.gov

20. **Selecting Colored Balls** Urn 1 contains 5 red balls and 3 black balls. Urn 2 contains 3 red balls and 1 black ball. Urn 3 contains 4 red balls and 2 black balls. If an urn is selected at random and a ball is drawn, find the probability it will be red.

21. **Automobile Insurance** An insurance company classifies drivers as low-risk, medium-risk, and high-risk. Of those insured, 60% are low-risk, 30% are medium-risk, and 10% are high-risk. After a study, the company finds that during a 1-year period, 1% of the low-risk drivers had an accident, 5% of the medium-risk drivers had an accident, and 9% of the high-risk drivers had an accident. If a driver is selected at random, find the probability that the driver will have had an accident during the year.

22. **Defective Items** A production process produces an item. On average, 15% of all items produced are defective. Each item is inspected before being shipped, and the inspector misclassifies an item 10% of the time. What proportion of the items will be "classified as good"? What is the probability that an item is defective given that it was classified as good?

23. **Prison Populations** For a recent year, 0.99 of the incarcerated population is adults and 0.07 of the incarcerated are adult females. If an incarcerated person is selected at random, find the probability that the person is a female given that the person is an adult.

    Source: Bureau of Justice

24. **Rolling Dice** Roll two standard dice and add the numbers. What is the probability of getting a number larger than 9 for the first time on the third roll?

25. **Heart Disease** Twenty-five percent of all deaths (all ages) are caused by diseases of the heart. Ischemic heart disease accounts for 16.4% of all deaths and heart failure for 2.3%. Choose one death at random. What is the probability that it is from ischemic heart disease given that it was from heart disease? Choose two deaths at random; what is the probability that at least one is from heart disease?

Source: Time Almanac.

26. **Country Club Activities** At the Avonlea Country Club, 73% of the members play bridge and swim, and 82% play bridge. If a member is selected at random, find the probability that the member swims, given that the member plays bridge.

27. **College Courses** At a large university, the probability that a student takes calculus and is on the dean's list is 0.042. The probability that a student is on the dean's list is 0.21. Find the probability that the student is taking calculus, given that he or she is on the dean's list.

28. **Congressional Terms** Below is given the summary from the 112th Congress of Senators whose terms end in 2013, 2015, or 2017.

	2013	2015	2017
Democrat	21	20	1
Republican	8	15	13

Choose one of these Senators at random and find

a. *P*(Democrat and term expires in 2015)

b. *P*(Republican or term expires in 2013)

c. *P*(Republican given term expires in 2017)

Are the events "Republican" and "term expires in 2015" independent? Explain.

Source: Time Almanac 2012.

29. **Pizza and Salads** In a pizza restaurant, 95% of the customers order pizza. If 65% of the customers order pizza and a salad, find the probability that a customer who orders pizza will also order a salad.

30. **Gift Baskets** The Gift Basket Store had the following premade gift baskets containing the following combinations in stock.

	Cookies	Mugs	Candy
Coffee	20	13	10
Tea	12	10	12

Choose 1 basket at random. Find the probability that it contains

a. Coffee or candy

b. Tea given that it contains mugs

c. Tea and cookies

Source: www.infoplease.com

31. **Blood Types and Rh Factors** In addition to being grouped into four types, human blood is grouped by its Rhesus (Rh) factor. Consider the figures below which show the distributions of these groups for Americans.

	O	A	B	AB
Rh+	37%	34%	10%	4%
Rh−	6	6	2	1

Choose one American at random. Find the probability that the person

a. Is a universal donor, i.e., has O-negative blood

b. Has type O blood given that the person is Rh+

c. Has A+ or AB− blood

d. Has Rh− given that the person has type B

Source: www.infoplease.com

32. **Doctor Specialties** Below are listed the numbers of doctors in various specialties by gender.

	Pathology	Pediatrics	Psychiatry
Male	12,575	33,020	27,803
Female	5,604	33,351	12,292

Choose one doctor at random.

a. Find *P* (male|pediatrician).

b. Find *P* (pathologist|female).

c. Are the characteristics "female" and "pathologist" independent? Explain.

Source: World Almanac.

33. **Lightning Strikes** It has been said that the probability of being struck by lightning is about 1 in 750,000, but under what circumstances? Below are listed the numbers of deaths from lightning since 1996.

	Golf/ball field	Boating/in water	Outside/camping	Construction	Under a tree	Phone	Other
1996–2000	16	23	117	9	40	0	30
2001–2005	17	16	112	3	35	0	23
2006–2010	15	17	91	0	42	1	16

Choose one fatality at random and find each probability.

a. Given that the death was after 2000, what is the probability that it occurred under a tree?

b. Find the probability that the death was from camping or being outside and was before 2001.

c. Find the probability that the death was from camping or being outside given that it was before 2001.

Source: Noaa.gov/lightstats

34. Foreign Adoptions The following foreign adoptions (in the United States) occurred during these particular years.

	2014	2013
China	2040	2306
Ethiopia	716	993
Ukraine	521	438

Choose one adoption at random from this chart.

a. What is the probability that it was from Ethiopia given that it was from 2013?

b. What is the probability that it was from the Ukraine and in 2014?

c. What is the probability that it did not occur in 2014 and was not from Ethiopia?

d. Choose two adoptions at random; what is the probability that they were both from China?

Source: World Almanac

35. Leisure Time Exercise Only 27% of U.S. adults get enough leisure time exercise to achieve cardiovascular fitness. Choose 3 adults at random. Find the probability that

a. All 3 get enough daily exercise

b. At least 1 of the 3 gets enough exercise

Source: www.infoplease.com

36. Customer Purchases In a department store there are 120 customers, 90 of whom will buy at least 1 item. If 5 customers are selected at random, one by one, find the probability that all will buy at least 1 item.

37. Marital Status of Women According to the *Statistical Abstract of the United States*, 70.3% of females ages 20 to 24 have never been married. Choose 5 young women in this age category at random. Find the probability that

a. None has ever been married

b. At least 1 has been married

Source: New York Times Almanac

38. Fatal Accidents The American Automobile Association (AAA) reports that of the fatal car and truck accidents, 54% are caused by car driver error. If 3 accidents are chosen at random, find the probability that

a. All are caused by car driver error

b. None is caused by car driver error

c. At least 1 is caused by car driver error

Source: AAA quoted on CNN

39. On-Time Airplane Arrivals According to FlightStats report released April 2013, the Salt Lake City airport led major U.S. airports in on-time arrivals with an 85.5% on-time rate. Choose 5 arrivals at random and find the probability that at least 1 was not on time.

Source: FlightStats

40. On-Time Flights A flight from Pittsburgh to Charlotte has a 90% on-time record. From Charlotte to Jacksonville, North Carolina, the flight is on time 80% of the time. The return flight from Jacksonville to Charlotte is on time 50% of the time and from Charlotte to Pittsburgh, 90% of the time. Consider a round trip from Pittsburgh to Jacksonville on these flights. Assume the flights are independent.

a. What is the probability that all 4 flights are on time?

b. What is the probability that at least 1 flight is not on time?

c. What is the probability that at least 1 flight is on time?

d. Which events are complementary?

41. Reading to Children Fifty-eight percent of American children (ages 3 to 5) are read to every day by someone at home. Suppose 5 children are randomly selected. What is the probability that at least 1 is read to every day by someone at home?

Source: Federal Interagency Forum on Child and Family Statistics.

42. Doctoral Assistantships Of Ph.D. students, 60% have paid assistantships. If 3 students are selected at random, find the probabilities that

a. All have assistantships

b. None has an assistantship

c. At least 1 has an assistantship

Source: U.S. Department of Education, Chronicle of Higher Education.

43. Drawing Cards If 5 cards are drawn at random from a deck of 52 cards and are not replaced, find the probability of getting at least one diamond.

44. Autism In recent years it was thought that approximately 1 in 110 children exhibited some form of autism. The most recent CDC study concluded that the proportion may be as high as 1 in 88. If indeed these new figures are correct, choose 3 children at random and find these probabilities.

a. What is the probability that none have autism?

b. What is the probability that at least 1 has autism?

c. Choose 10 children at random. What is the probability that at least 1 has autism?

Source: cdc.gov

45. Video Game. Video games are rated according to the content. The average age of a gamer is 35 years old. In 2015, 15.5% of the video games were rated Mature. Choose 5 purchased games at random. Find the probability that

a. None of the five were rated mature.

b. At least 1 of the 5 was rated mature.

Source: Nobullying.com

46. Medication Effectiveness A medication is 75% effective against a bacterial infection. Find the probability that if 12 people take the medication, at least 1 person's infection will not improve.

47. **Selecting Digits** If 3 digits are selected at random with replacement, find the probability of getting at least one odd number. Would you consider this event likely or unlikely? Why?

48. **Selecting a Letter of the Alphabet** If 3 letters of the alphabet are selected at random, find the probability of getting at least 1 letter x. Letters can be used more than once. Would you consider this event likely to happen? Explain your answer.

49. **Rolling a Die** A die is rolled twice. Find the probability of getting at least one 6.

50. **U.S. Organ Transplants** As of June 2015, 81.4% of patients were waiting on a kidney, 11.7% were waiting on a liver, and 3.1% were waiting on a heart. Choose 6 patients on the transplant waiting list at random in 2015. Find the probability that
    a. All were waiting for a kidney.
    b. None were waiting for a kidney.
    c. At least 1 was waiting for a kidney.

51. **Lucky People** Twelve percent of people in Western countries consider themselves lucky. If 3 people are selected at random, what is the probability that at least one will consider himself lucky?
    Source: San Diego Tribune

52. **Selecting a Flower** In a large vase, there are 8 roses, 5 daisies, 12 lilies, and 9 orchids. If 4 flowers are selected at random, and not replaced, find the probability that at least 1 of the flowers is a rose. Would you consider this event likely to occur? Explain your answer.

Extending the Concepts

53. Let A and B be two mutually exclusive events. Are A and B independent events? Explain your answer.

54. **Types of Vehicles** The Bargain Auto Mall has the following cars in stock.

	SUV	Compact	Mid-sized
Foreign	20	50	20
Domestic	65	100	45

    Are the events "compact" and "domestic" independent? Explain.

55. **College Enrollment** An admissions director knows that the probability a student will enroll after a campus visit is 0.55, or $P(E) = 0.55$. While students are on campus visits, interviews with professors are arranged. The admissions director computes these conditional probabilities for students enrolling after visiting three professors, DW, LP, and MH.

$P(E|\text{DW}) = 0.95$ $P(E|\text{LP}) = 0.55$ $P(E|\text{MH}) = 0.15$

    Is there something wrong with the numbers? Explain.

56. **Commercials** Event A is the event that a person remembers a certain product commercial. Event B is the event that a person buys the product. If $P(B) = 0.35$, comment on each of these conditional probabilities if you were vice president for sales.
    a. $P(B|A) = 0.20$
    b. $P(B|A) = 0.35$
    c. $P(B|A) = 0.55$

57. Given a sample space with events A and B such that $P(A) = 0.342$, $P(B) = 0.279$, and $P(A \text{ or } B) = 0.601$. Are A and B mutually exclusive? Are A and B independent? Find $P(A|B)$, $P(\text{not } B)$, and $P(A \text{ and } B)$.

58. **Child's Board Game** In a child's board game of the tortoise and the hare, the hare moves by roll of a standard die and the tortoise by a six-sided die with the numbers 1, 1, 1, 2, 2, and 3. Roll each die once. What is the probability that the tortoise moves ahead of the hare?

59. **Bags Containing Marbles** Two bags contain marbles. Bag 1 contains 1 black marble and 9 white marbles. Bag 2 contains 1 black marble and x white marbles. If you choose a bag at random, then choose a marble at random, the probability of getting a black marble is $\frac{1}{3}$. How many white marbles are in bag 2? ■

4–4 Counting Rules

Many times a person must know the number of all possible outcomes for a sequence of events. To determine this number, three rules can be used: the *fundamental counting rule*, the *permutation rule*, and the *combination rule*. These rules are explained here, and they will be used in Section 4–5 to find probabilities of events.

The first rule is called the **fundamental counting rule.**

The Fundamental Counting Rule

OBJECTIVE 5

Find the total number of outcomes in a sequence of events, using the fundamental counting rule

Fundamental Counting Rule

In a sequence of n events in which the first one has k_1 possibilities and the second event has k_2 and the third has k_3, and so forth, the total number of possibilities of the sequence will be

$$k_1 \cdot k_2 \cdot k_3 \cdots k_n$$

***Note:* In this case *and* means to multiply.**

Examples 4–38 through 4–41 illustrate the fundamental counting rule.

EXAMPLE 4–38 Tossing a Coin and Rolling a Die

A coin is tossed and a die is rolled. Find the number of outcomes for the sequence of events.

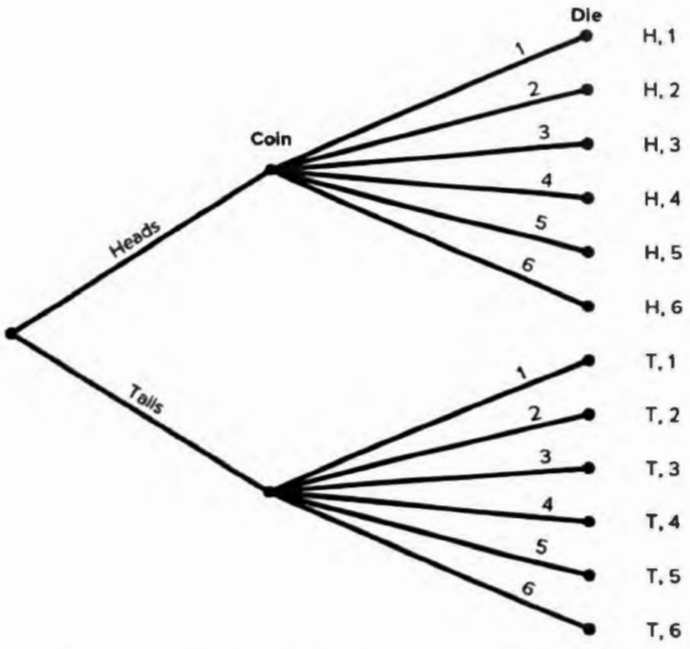
Interesting Fact

Possible games of chess: 25×10^{76}

SOLUTION

Since the coin can land either heads up or tails up and since the die can land with any one of six numbers showing face up, there are $2 \cdot 6 = 12$ possibilities. A tree diagram can also be drawn for the sequence of events. See Figure 4–10.

FIGURE 4–10
Complete Tree Diagram for Example 4–38



EXAMPLE 4–39 Types of Paint

A paint manufacturer wishes to manufacture several different paints. The categories include

Color	Red, blue, white, black, green, brown, yellow
Type	Latex, oil
Texture	Flat, semigloss, high gloss
Use	Outdoor, indoor

How many different kinds of paint can be made if you can select one color, one type, one texture, and one use?

SOLUTION

You can choose one color and one type and one texture and one use. Since there are 7 color choices, 2 type choices, 3 texture choices, and 2 use choices, the total number of possible different paints is as follows:

Color		Type		Texture		Use	
7	·	2	·	3	·	2	= 84

EXAMPLE 4–40 Distribution of Blood Types

There are four blood types, A, B, AB, and O. Blood can also be Rh+ and Rh−. Finally, a blood donor can be classified as either male or female. How many different ways can a donor have his or her blood labeled?

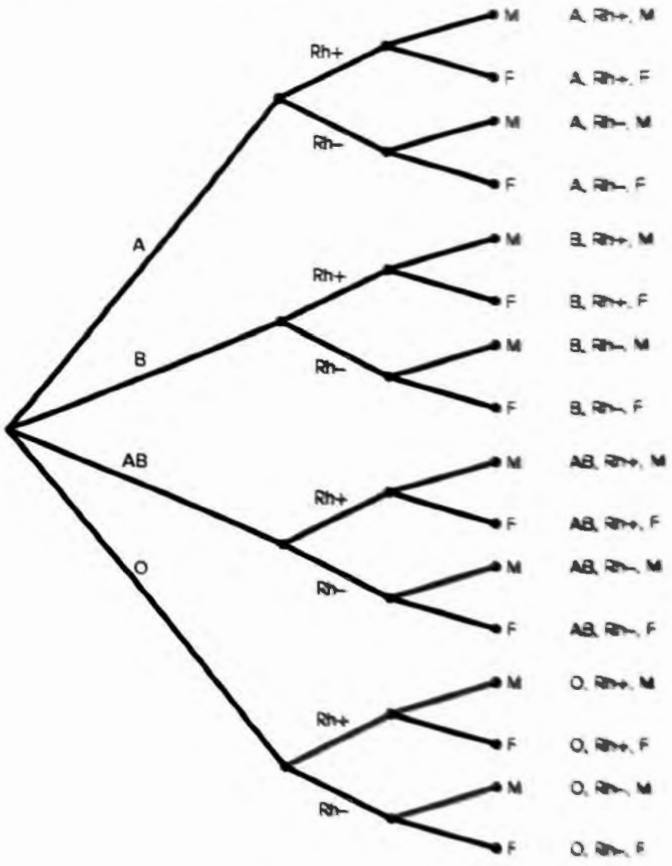
SOLUTION

Since there are 4 possibilities for blood type, 2 possibilities for Rh factor, and 2 possibilities for the gender of the donor, there are 4 · 2 · 2, or 16, different classification categories, as shown.

Blood type		Rh		Gender	
4	·	2	·	2	= 16

A tree diagram for the events is shown in Figure 4–11.

FIGURE 4–11
Complete Tree Diagram for Example 4–40



When determining the number of different possibilities of a sequence of events, you must know whether repetitions are permissible.

EXAMPLE 4–41 Railroad Memorial License Plates

The first year the state of Pennsylvania issued railroad memorial license plates, the plates had a picture of a steam engine followed by four digits. Assuming that repetitions are allowed, how many railroad memorial plates could be issued?

SOLUTION

Since there are four spaces to fill for each space, the total number of plates that can be issued is $10 \cdot 10 \cdot 10 \cdot 10 = 10{,}000$. *Note:* Actually there was such a demand for the plates, Pennsylvania had to use letters also.

Now if repetitions are not permitted, the first digit in the plate in Example 4–41 could be selected in 10 ways, the second digit in 9 ways, etc. So the total number of plates that could be issued is $10 \cdot 9 \cdot 8 \cdot 7 = 5040$.

The same situation occurs when one is drawing balls from an urn or cards from a deck. If the ball or card is replaced before the next one is selected, then repetitions are permitted, since the same one can be selected again. But if the selected ball or card is not replaced, then repetitions are not permitted, since the same ball or card cannot be selected the second time.

These examples illustrate the fundamental counting rule. In summary: *If repetitions are permitted, then the numbers stay the same going from left to right. If repetitions are not permitted, then the numbers decrease by* 1 *for each place left to right.*

Two other rules that can be used to determine the total number of possibilities of a sequence of events are the permutation rule and the combination rule.

Factorial Notation

Historical Note

In 1808 Christian Kramp first used factorial notation.

These rules use *factorial notation.* The factorial notation uses the exclamation point.

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$
$$9! = 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$

To use the formulas in the permutation and combination rules, a special definition of 0! is needed: $0! = 1$.

Factorial Formulas

For any counting number n

$$n! = n(n-1)(n-2)\cdots 1$$
$$0! = 1$$

Permutations

A permutation is an arrangement of n objects in a specific order.

Examples 4–42 and 4–43 illustrate permutations.

EXAMPLE 4–42 Lock Codes

A person needs to make up a four-digit combination to open the lock on the garage door. The person decides to use four different digits from the first four digits, 0 through 3. How many different combinations can be made?

SOLUTION

The person has four choices for the first digit, three ways for the second digit, etc. Hence,

$$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

There are 24 different combinations that can be used.

In Example 4–42 all objects were used up. But what happens when not all objects are used up? The answer to this question is given in Example 4–43.

EXAMPLE 4–43 Business Location

A business owner wishes to rank the top 3 locations selected from 5 locations for a business. How many different ways can she rank them?

SOLUTION

Using the fundamental counting rule, she can select any one of the 5 for first choice, then any one of the remaining 4 locations for her second choice, and finally, any one of the remaining locations for her third choice, as shown.

First choice		Second choice		Third choice	
5	·	4	·	3	= 60

The solutions in Examples 4–42 and 4–43 are permutations.

OBJECTIVE 6

Find the number of ways that r objects can be selected from n objects, using the permutation rule.

Permutation Rule 1

The arrangement of n objects in a specific order using r objects at a time is called a *permutation of n objects taking r objects at a time*. It is written as ${}_nP_r$, and the formula is

$${}_nP_r = \frac{n!}{(n-r)!}$$

The notation ${}_nP_r$ is used for permutations.

$${}_6P_4 \text{ means } \frac{6!}{(6-4)!} \quad \text{or} \quad \frac{6!}{2!} = \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = 360$$

Although Examples 4–42 and 4–43 were solved by the multiplication rule, they can now be solved by the permutation rule.

In Example 4–42, five locations were taken and then arranged in order; hence,

$${}_5P_5 = \frac{5!}{(5-5)!} = \frac{5!}{0!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{1} = 120$$

(Recall that $0! = 1$.)

In Example 4–43, three locations were selected from 5 locations, so $n = 5$ and $r = 3$; hence,

$${}_5P_3 = \frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = 60$$

Examples 4–44 and 4–45 illustrate the permutation rule.

EXAMPLE 4–44 Radio Show Guests

A radio talk show host can select 3 of 6 special guests for her program. The order of appearance of the guests is important. How many different ways can this be done?

SOLUTION

Since the order of appearance on the show is important, there are ${}_6P_3$ ways to select the guests.

$$ {}_6P_3 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = \frac{6 \cdot 5 \cdot 4 \cdot 3!}{3!} = 120 $$

Hence, there can be 120 different ways to select 3 guests and present them on the program in a specific order.

EXAMPLE 4–45 School Musical Plays

A school musical director can select 2 musical plays to present next year. One will be presented in the fall, and one will be presented in the spring. If she has 9 to pick from, how many different possibilities are there?

SOLUTION

Order is important since one play can be presented in the fall and the other play in the spring.

$$ {}_9P_2 = \frac{9!}{(9-2)!} = \frac{9!}{7!} = \frac{9 \cdot 8 \cdot 7!}{7!} = 72 $$

There are 72 different possibilities.

In the previous examples, all items involving permutations were different, but when some of the items are identical, a second permutation rule can be used.

Permutation Rule 2

The number of permutations of n objects when r_1 objects are identical, r_2 objects are identical, . . . , r_p objects are identical, etc., is

$$ \frac{n!}{r_1!r_2! \cdots r_p!} $$

where $r_1 + r_2 + \cdots + r_p = n$.

EXAMPLE 4–46 Letter Permutations

How many permutations of the letters can be made from the word *STATISTICS*?

SOLUTION

In the word *STATISTICS*, there are 3 S's, 3 T's, 2 I's, 1 A, and 1 C.

$$ \frac{10!}{3!3!2!1!1!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 3 \cdot 2 \cdot 1 \cdot 2 \cdot 1 \cdot 1 \cdot 1} = 50{,}400 $$

There are 50,400 permutations that can be made from the word *STATISTICS*.

Combinations

Suppose a dress designer wishes to select two colors of material to design a new dress, and she has on hand four colors. How many different possibilities can there be in this situation?

OBJECTIVE 7

Find the number of ways that r objects can be selected from n objects without regard to order, using the combination rule.

This type of problem differs from previous ones in that the order of selection is not important. That is, if the designer selects yellow and red, this selection is the same as the selection red and yellow. This type of selection is called a *combination*. The difference between a permutation and a combination is that in a combination, the order or arrangement of the objects is not important; by contrast, order *is* important in a permutation. Example 4–46 illustrates this difference.

A selection of distinct objects without regard to order is called a combination.

© The McGraw-Hill Companies, Inc./Evelyn Jo Hebert, photographer

The difference between a combination and a permutation can be shown using the letters A, B, C, and D. The permutations for the letters A, B, C, and D are

AB	BA	CA	DA
AC	BC	CB	DB
AD	BD	CD	DC

In permutations, AB is different from BA. But in combinations, AB is the same as BA since the order of the objects does not matter in combinations. Therefore, if duplicates are removed from a list of permutations, what is left is a list of combinations, as shown.

AB	BA	CA	DA
AC	BC	CB	DB
AD	BD	CD	DC

Hence, the combinations of A, B, C, and D are AB, AC, AD, BC, BD, and CD. (Alternatively, BA could be listed and AB crossed out, etc.) The combinations have been listed alphabetically for convenience, but this is not a requirement.

Interesting Fact

The total number of hours spent moving boxes in the United States each year is 2,220,000,000.

Combinations are used when the order or arrangement is not important, as in the selecting process. Suppose a committee of 5 students is to be selected from 25 students. The 5 selected students represent a combination, since it does not matter who is selected first, second, etc.

Combination Rule

The number of combinations of r objects selected from n objects is denoted by ${}_nC_r$ and is given by the formula

$$ {}_nC_r = \frac{n!}{(n-r)!r!} $$

EXAMPLE 4–47 Combinations

How many combinations of 4 objects are there, taken 2 at a time?

SOLUTION

Since this is a combination problem, the answer is

$$ {}_4C_2 = \frac{4!}{(4-2)!2!} = \frac{4!}{2!2!} = \frac{\overset{2}{\cancel{4}} \cdot 3 \cdot \cancel{2}!}{\cancel{2} \cdot 1 \cdot \cancel{2}!} = 6 $$

This is the same result shown on the previous page.

Notice that the expression for ${}_nC_r$ is

$$ \frac{n!}{(n-r)!r!} $$

which is the formula for permutations with $r!$ in the denominator. In other words,

$$ {}_nC_r = \frac{{}_nP_r}{r!} $$

This $r!$ divides out the duplicates from the number of permutations. For each two letters, there are two permutations but only one combination. Hence, dividing the number of permutations by $r!$ eliminates the duplicates. This result can be verified for other values of n and r. Note ${}_nC_n = 1$.

EXAMPLE 4–48 Movies at the Park

The director of Movies at the Park must select 4 movies from a total of 10 movies to show on Movie Night at the Park. How many different ways can the selections be made?

SOLUTION

$$ {}_{10}C_4 = \frac{10!}{(10-4)!4!} = \frac{10!}{6!4!} = \frac{10}{4} \cdot \frac{9}{3} \cdot \frac{8}{2} \cdot \frac{7}{1} = 210 $$

The director has 210 different ways to select four movies from 10 movies. In this case, the order in which the movies are shown is not important.

EXAMPLE 4–49 Committee Selection

In a club there are 7 women and 5 men. A committee of 3 women and 2 men is to be chosen. How many different possibilities are there?

SOLUTION

Here, you must select 3 women from 7 women, which can be done in $_7C_3$ or 35 ways. Next, 2 men must be selected from 5 men, which can be done in $_5C_2$ or 10 ways. Finally, by the fundamental counting rule, the total number of different ways is $35 \cdot 10 = 350$, since you are choosing both men and women. Using the formula gives

$$_7C_3 \cdot {}_5C_2 = \frac{7!}{(7-3)!3!} \cdot \frac{5!}{(5-2)!2!} = 350$$

Table 4–1 summarizes the counting rules.

TABLE 4–1 Summary of Counting Rules

Rule	Definition	Formula
Fundamental counting rule	The number of ways a sequence of n events can occur if the first event can occur in k_1 ways, the second event can occur in k_2 ways, etc.	$k_1 \cdot k_2 \cdot k_3 \cdot \ldots$
Permutation rule 1	The number of permutations of n objects taking r objects at a time (order is important)	$_nP_r = \frac{n!}{(n-r)!}$
Permutation rule 2	The number of permutations of n objects when n_1 objects are identical, n_2 objects are identical, ..., n_k objects are identical	$\frac{n!}{n_1!\,n_2!\,\ldots\,n_k!}$
Combination rule	The number of combinations of r objects taken from n objects (order is not important)	$_nC_r = \frac{n!}{(n-r)!r!}$

Applying the Concepts 4–4

Garage Door Openers

Garage door openers originally had a series of four on/off switches so that homeowners could personalize the frequencies that opened their garage doors. If all garage door openers were set at the same frequency, anyone with a garage door opener could open anyone else's garage door.

1. Use a tree diagram to show how many different positions 4 consecutive on/off switches could be in.

After garage door openers became more popular, another set of 4 on/off switches was added to the systems.

2. Find a pattern of how many different positions are possible with the addition of each on/off switch.
3. How many different positions are possible with 8 consecutive on/off switches?
4. Is it reasonable to assume, if you owned a garage door opener with 8 switches, that someone could use his or her garage door opener to open your garage door by trying all the different possible positions?

For a specific year it was reported that the ignition keys for Dodge Caravans were made from a single blank that had five cuts on it. Each cut was made at one out of five possible levels. For that year assume there were 420,000 Dodge Caravans sold in the United States.

5. How many different possible keys can be made from the same key blank?
6. How many different Dodge Caravans could any one key start?

Look at the ignition key for your car and count the number of cuts on it. Assume that the cuts are made at one of any of five possible levels. Most car companies use one key blank for all their makes and models of cars.

7. Conjecture how many cars your car company sold over recent years, and then figure out how many other cars your car key could start. What would you do to decrease the odds of someone being able to open another vehicle with his or her key?

See pages 254–255 for the answers.

Exercises 4–4

1. **Zip Codes** How many 5-digit zip codes are possible if digits can be repeated? If there cannot be repetitions?
2. **Letter Permutations** List all the permutations of the letters in the word *MATH*.
3. **Speaking Order** Seven elementary students are selected to give a 3-minute presentation on what they did during summer vacation. How many different ways can the speakers be arranged?
4. **Visiting Nurses** How many different ways can a visiting nurse visit 9 patients if she wants to visit them all in one day?
5. **Quinto Lottery** A lottery game called Quinto is played by choosing five numbers each, from 0 through 9. How many numbers are possible? Although repeats are allowed, how many numbers are possible if repeats are not allowed?
6. **Show Programs** Three bands and two comics are performing for a student talent show. How many different programs (in terms of order) can be arranged? How many if the comics must perform between bands?
7. **Rolling Dice** If five dice are rolled, how many different outcomes are there?
8. **Radio Station Call Letters** The call letters of a radio station must have 4 letters. The first letter must be a K or a W. How many different station call letters can be made if repetitions are not allowed? If repetitions are allowed?
9. **Film Showings** At the Rogue Film Festival, the director must select one film from each category. There are 6 drama films, 3 sci-fi films, and 5 comedy films. How many different ways can a film be selected?
10. **Secret Code Word** How many 4-letter code words can be made using the letters in the word *pencil* if repetitions are permitted? If repetitions are not permitted?
11. **Passwords** Given the characters *A, B, C, H, I, T, U, V,* 1, 2, 3, and 4, how many seven-character passwords can be made? (No repeats are allowed.) How many if you have to use all four numbers as the first four characters in the password?
12. **Automobile Trips** There are 2 major roads from city *X* to city *Y* and 4 major roads from city *Y* to city *Z*. How many different trips can be made from city *X* to city *Z*, passing through city *Y*?
13. Evaluate each expression.

a. 11!	*e.* ${}_6P_4$	*i.* ${}_9P_2$
b. 9!	*f.* ${}_{12}P_8$	*j.* ${}_{11}P_3$
c. 0!	*g.* ${}_7P_7$	
d. 1!	*h.* ${}_4P_0$	

14. Evaluate each expression.

a. 6!	*e.* ${}_9P_6$
b. 11!	*f.* ${}_{11}P_4$
c. 2!	*g.* ${}_8P_0$
d. 9!	*h.* ${}_{10}P_2$

15. **Sports Car Stripes** How many different 4-color code stripes can be made on a sports car if each code consists of the colors green, red, blue, and white? All colors are used only once.
16. **Manufacturing Tests** An inspector must select 3 tests to perform in a certain order on a manufactured part. He has a choice of 7 tests. How many ways can he perform 3 different tests?
17. **Endangered Amphibians** There are 9 endangered amphibian species in the United States. How many ways

can a student select 3 of these species to write a report about them? The order of selection is important.

18. **Inspecting Restaurants** How many different ways can a city health department inspector visit 5 restaurants in a city with 10 restaurants?

19. **Word Permutation** How many different 4-letter permutations can be written from the word *hexagon*?

20. **Cell Phone Models** A particular cell phone company offers 4 models of phones, each in 6 different colors and each available with any one of 5 calling plans. How many combinations are possible?

21. **ID Cards** How many different ID cards can be made if there are 6 digits on a card and no digit can be used more than once?

22. **Free-Sample Requests** An online coupon service has 13 offers for free samples. How many different requests are possible if a customer must request exactly 3 free samples? How many are possible if the customer may request up to 3 free samples?

23. **Ticket Selection** How many different ways can 4 tickets be selected from 50 tickets if each ticket wins a different prize?

24. **Movie Selections** The Foreign Language Club is showing a four-movie marathon of subtitled movies. How many ways can they choose 4 from the 11 available?

25. **Task Assignments** How many ways can an adviser choose 4 students from a class of 12 if they are all assigned the same task? How many ways can the students be chosen if they are each given a different task?

26. **Agency Cases** An investigative agency has 7 cases and 5 agents. How many different ways can the cases be assigned if only 1 case is assigned to each agent?

27. **Signal Flags** How many different flag signals, each consisting of 7 flags hung vertically, can be made when there are 3 indistinguishable red flags, 2 blue flags, and 2 white flags?

28. **Word Permutations** How many permutations can be made using all the letters in the word *MASSACHUSETTS*?

29. **Code Words** How many different 9-letter code words can be made using the symbols %, %, %, %, &, &, &, +, +?

30. **Toothpaste Display** How many different ways can 5 identical tubes of tartar control toothpaste, 3 identical tubes of bright white toothpaste, and 4 identical tubes of mint toothpaste be arranged in a grocery store counter display?

31. **Book Arrangements** How many different ways can 6 identical hardback books, 3 identical paperback books, and 3 identical boxed books be arranged on a shelf in a bookstore?

32. **Letter Permutations** How many different permutations of the letters in the word *CINCINNATI* are there?

33. Evaluate each expression.
   a. ${}_4C_2$
   b. ${}_4C_3$
   c. ${}_7C_4$
   d. ${}_6C_$
   e. ${}_4C_4$

34. Evaluate each expression.
   a. ${}_3C_0$
   b. ${}_3C_3$
   c. ${}_9C_7$
   d. ${}_{12}C_2$
   e. ${}_4C_3$

35. **Medications for Depression** A researcher wishes her patients to try a new medicine for depression. How many different ways can she select 5 patients from 50 patients?

36. **Selecting Players** How many ways can 4 baseball players and 3 basketball players be selected from 12 baseball players and 9 basketball players?

37. **Coffee Selection** A coffee shop serves 12 different kinds of coffee drinks. How many ways can 4 different coffee drinks be selected?

38. **Selecting Christmas Presents** If a person can select 3 presents from 10 presents under a Christmas tree, how many different combinations are there?

39. **Buffet Desserts** In how many ways can you choose 3 kinds of ice cream and 2 toppings from a dessert buffet with 10 kinds of ice cream and 6 kinds of toppings?

40. **Bridge Foursomes** How many different tables of 4 can you make from 16 potential bridge players? How many different tables if 4 of the players insist on playing together?

41. **Music Recital** Six students are performing one song each in a jazz vocal recital. Two students have repertoires of five numbers, and the others have four songs each prepared. How many different programs are possible without regard to order? Assume that the repertory selections are all unique.

42. **Freight Train Cars** In a train yard there are 4 tank cars, 12 boxcars, and 7 flatcars. How many ways can a train be made up consisting of 2 tank cars, 5 boxcars, and 3 flatcars? (In this case, order is not important.)

43. **Selecting a Committee** There are 7 women and 5 men in a department. How many ways can a committee of 4 people be selected? How many ways can this committee be selected if there must be 2 men and 2 women on the committee? How many ways can this committee be selected if there must be at least 2 women on the committee?

44. **Selecting Cereal Boxes** Wake Up cereal comes in 2 types, crispy and crunchy. If a researcher has 10 boxes of each, how many ways can she select 3 boxes of each for a quality control test?

45. **Hawaiian Words** The Hawaiian alphabet consists of 7 consonants and 5 vowels. How many three-letter "words" are possible if there are never two consonants together and if a word must always end in a vowel?

46. **Selecting a Jury** How many ways can a jury of 6 women and 6 men be selected from 10 women and 12 men?

47. **Selecting Students** How many ways can you pick 4 students from 10 students (6 men, 4 women) if you must have an equal number of each gender or all of the same gender?

48. **Investigative Team** The state narcotics bureau must form a 5-member investigative team. If it has 25 agents from which to choose, how many different possible teams can be formed?

49. **Dominoes** A domino is a flat rectangular block whose face is divided into two square parts, each part showing from zero to six pips (or dots). Playing a game consists of playing dominoes with a matching number of pips. Explain why there are 28 dominoes in a complete set.

50. **Charity Event Participants** There are 16 seniors and 15 juniors in a particular social organization. In how many ways can 4 seniors and 2 juniors be chosen to participate in a charity event?

51. **Automobile Selection** An automobile dealer has 12 small automobiles, 8 mid-size automobiles, and 6 large automobiles on his lot. How many ways can two of each type of automobile be selected from his inventory?

52. **DVD Selection** How many ways can a person select 8 DVDs from a display of 13 DVDs?

53. **Railroad Accidents** A researcher wishes to study railroad accidents. He wishes to select 3 railroads from 10 Class I railroads, 2 railroads from 6 Class II railroads, and 1 railroad from 5 Class III railroads. How many different possibilities are there for his study?

54. **Selecting a Location** An advertising manager decides to have an ad campaign in which 8 special calculators will be hidden at various locations in a shopping mall. If he has 17 locations from which to pick, how many different possible combinations can he choose?

Permutations and Combinations

55. **Selecting Posters** A buyer decides to stock 8 different posters. How many ways can she select these 8 if there are 20 from which to choose?

56. **Test Marketing Products** Anderson Research Company decides to test-market a product in 6 areas. How many different ways can 3 areas be selected in a certain order for the first test?

57. **Nuclear Power Plants** How many different ways can a government researcher select 5 nuclear power plants from 9 nuclear power plants in Pennsylvania?

58. **Selecting Musicals** How many different ways can a theatrical group select 2 musicals and 3 dramas from 11 musicals and 8 dramas to be presented during the year?

59. **Textbook Selection** How many different ways can an instructor select 2 textbooks from a possible 17?

60. **DVD Selection** How many ways can a person select 8 DVDs from 10 DVDs?

61. **Flight Attendants** How many different ways can 3 flight attendants be selected from 11 flight attendants for a routine flight?

62. **Signal Flags** How many different signals can be made by using at least 3 different flags if there are 5 different flags from which to select?

63. **Dinner Selections** How many ways can a dinner patron select 3 appetizers and 2 vegetables if there are 6 appetizers and 5 vegetables on the menu?

64. **Air Pollution** The Environmental Protection Agency must investigate 9 mills for complaints of air pollution. How many different ways can a representative select 5 of these to investigate this week?

65. **Selecting Officers** In a board of directors composed of 8 people, how many ways can one chief executive officer, one director, and one treasurer be selected?

66. **Selecting Council Members** The presidents, vice presidents, and secretary-treasurers from each of four classes are eligible for an all-school council. How many ways can four officers be chosen from these representatives? How many ways can they be chosen if the president must be selected from the sitting presidents, the vice president from the sitting vice presidents, the secretary from the sitting secretary-treasurers, and the treasurer from everybody who's left?

Extending the Concepts

67. **Selecting Coins** How many different ways can you select one or more coins if you have 2 nickels, 1 dime, and 1 half-dollar?

68. **People Seated in a Circle** In how many ways can 3 people be seated in a circle? 4? n? (*Hint:* Think of them standing in a line before they sit down and/or draw diagrams.)

69. **Seating in a Movie Theater** How many different ways can 5 people—A, B, C, D, and E—sit in a row at a movie theater if (*a*) A and B must sit together; (*b*) C must sit to the right of, but not necessarily next to, B; (*c*) D and E will not sit next to each other?

70. **Poker Hands** Using combinations, calculate the number of each type of poker hand in a deck of

4–5 Probability and Counting Rules

OBJECTIVE 8

Find the probability of an event, using the counting rules.

The counting rules can be combined with the probability rules in this chapter to solve many types of probability problems. By using the fundamental counting rule, the permutation rules, and the combination rule, you can compute the probability of outcomes of many experiments, such as getting a full house when 5 cards are dealt or selecting a committee of 3 women and 2 men from a club consisting of 10 women and 10 men.

© Royalty-Free/CORBIS

EXAMPLE 4–50 Four Aces

Find the probability of getting 4 aces when 5 cards are drawn from an ordinary deck of cards.

SOLUTION

There are ${}_{52}C_5$ ways to draw 5 cards from a deck. There is only 1 way to get 4 aces (that is, ${}_4C_4$), but there are 48 possibilities to get the fifth card. Therefore, there are 48 ways to get 4 aces and 1 other card. Hence,

$$P(\text{4 aces}) = \frac{{}_4C_4 \cdot 48}{{}_{52}C_5} = \frac{1 \cdot 48}{2{,}598{,}960} = \frac{48}{2{,}598{,}960} = \frac{1}{54{,}145}$$

EXAMPLE 4–51 Defective Integrated Circuits

A box contains 24 integrated circuits, 4 of which are defective. If 4 are sold at random, find the following probabilities.

a. Exactly 2 are defective.

b. None is defective.

c. All are defective.

d. At least 1 is defective.

SOLUTION

There are ${}_{24}C_4$ ways to sell 4 integrated circuits, so the denominator in each case will be 10,626.

a. Two defective integrated circuits can be selected as ${}_4C_2$ and two nondefective ones as ${}_{20}C_2$. Hence,

$$P(\text{exactly 2 defectives}) = \frac{{}_4C_2 \cdot {}_{20}C_2}{{}_{24}C_4} = \frac{1140}{10{,}626} = \frac{190}{1771}$$

b. The number of ways to choose no defectives is ${}_{20}C_4$. Hence,

$$P(\text{no defectives}) = \frac{{}_{20}C_4}{{}_{24}C_4} = \frac{4845}{10{,}626} = \frac{1615}{3542}$$

c. The number of ways to choose 4 defectives from 4 is ${}_4C_4$, or 1. Hence,

$$P(\text{all defective}) = \frac{1}{{}_{24}C_4} = \frac{1}{10{,}626}$$

d. To find the probability of at least 1 defective transistor, find the probability that there are no defective integrated circuits, and then subtract that probability from 1.

$$P(\text{at least 1 defective}) = 1 - P(\text{no defectives})$$
$$= 1 - \frac{{}_{20}C_4}{{}_{24}C_4} = 1 - \frac{1615}{3542} = \frac{1927}{3542}$$

EXAMPLE 4–52 Term Paper Selection

A student needs to select two topics to write two term papers for a course. There are 8 topics in economics and 11 topics in science. Find the probability that she selects one topic in economics and one topic is science to complete her assignment.

SOLUTION

$$P(\text{economics and science}) = \frac{{}_8C_1 \cdot {}_{11}C_1}{{}_{19}C_2}$$
$$= \frac{88}{171} = 0.515 \approx 51.5\%$$

Hence, there is about a 51.5% probability that a person will select one topic from economics and one topic from science.

EXAMPLE 4–53 State Lottery Number

In the Pennsylvania State Lottery, a person selects a three-digit number and repetitions are permitted. If a winning number is selected, find the probability that it will have all three digits the same.

SOLUTION

Since there are 10 different digits, there are $10 \cdot 10 \cdot 10 = 1000$ ways to select a winning number. When all of the digits are the same, that is, 000, 111, 222, . . . , 999, there are 10 possibilities, so the probability of selecting a winning number which has 3 identical digits is $\frac{10}{1000} = \frac{1}{100}$.

EXAMPLE 4–54 Tennis Tournament

There are 8 married couples in a tennis club. If 1 man and 1 woman are selected at random to plan the summer tournament, find the probability that they are married to each other.

SOLUTION

Since there are 8 ways to select the man and 8 ways to select the woman, there are $8 \cdot 8$, or 64, ways to select 1 man and 1 woman. Since there are 8 married couples, the solution is $\frac{8}{64} = \frac{1}{8}$.

As indicated at the beginning of this section, the counting rules and the probability rules can be used to solve a large variety of probability problems found in business, gambling, economics, biology, and other fields.

SPEAKING OF STATISTICS

The Mathematics of Gambling

Gambling is big business. There are state lotteries, casinos, sports betting, and church bingos. It seems that today everybody is either watching or playing Texas Hold 'Em Poker.

Using permutations, combinations, and the probability rules, mathematicians can find the probabilities of various gambling games. Here are the probabilities of the various 5-card poker hands.

Hand	Number of ways	Probability
Straight flush	40	0.000015
Four of a kind	624	0.000240
Full house	3,744	0.001441
Flush	5,108	0.001965
Straight	10,200	0.003925
Three of a kind	54,912	0.021129
Two pairs	123,552	0.047539
One pair	1,098,240	0.422569
Less than one pair	1,302,540	0.501177
Total	2,598,960	1.000000

© Royalty-Free/CORBIS

The chance of winning at gambling games can be compared by using what is called the house advantage, house edge, or house percentage. For example, the house advantage for roulette is about 5.26%, which means in the long run, the house wins 5.26 cents on every $1 bet; or you will lose, on average, 5.26 cents on every $1 you bet. The lower the house advantage, the more favorable the game is to you.

For the game of craps, the house advantage is anywhere between 1.4 and 15%, depending on what you bet on. For the game called Keno, the house advantage is 29.5%. The house advantage for Chuck-a-Luck is 7.87%, and for Baccarat, it is either 1.36 or 1.17%, depending on your bet.

Slot machines have a house advantage anywhere from about 4 to 10% depending on the geographic location, such as Atlantic City, Las Vegas, and Mississippi, and the amount put in the machine, such as 5 cents, 25 cents, and $1.

Actually, gamblers found winning strategies for the game blackjack or 21, such as card counting. However, the casinos retaliated by using multiple decks and by banning card counters.

Applying the Concepts 4–5

Counting Rules and Probability

One of the biggest problems for students when doing probability problems is to decide which formula or formulas to use. Another problem is to decide whether two events are independent or dependent. Use the following problem to help develop a better understanding of these concepts.

Assume you are given a five-question multiple-choice quiz. Each question has 5 possible answers: A, B, C, D, and E.

1. How many events are there?
2. Are the events independent or dependent?
3. If you guess at each question, what is the probability that you get all of them correct?
4. What is the probability that a person guesses answer A for each question?

Assume that you are given a five-question matching test in which you are to match the correct answers in the right column with the questions in the left column. You can use each answer only once.

5. How many events are there?
6. Are the events independent or dependent?
7. What is the probability of getting them all correct if you are guessing?
8. What is the difference between the two problems?

See pages 254–255 for the answers.

Exercises 4–5

1. **Selecting Cards** Find the probability of getting 2 face cards (king, queen, or jack) when 2 cards are drawn from a deck without replacement.
2. **Selecting Cards** Cards numbered 1–10 are shuffled and dealt face down. What is the probability that they are in order?
3. **Educational Fellowship** A university received 9 applications for three postdoctorate fellowships. Five of the applicants are men and four are women. Find these probabilities.
   a. All 3 who are selected are men.
   b. All 3 who are selected are women.
   c. Two men and one woman are selected.
   d. Two women and one man are selected.
4. **Senate Partisanship** The composition of the Senate of the 114th Congress is

   54 Republicans 2 Independent 44 Democrats

   A new committee is being formed to study ways to benefit the arts in education. If 3 Senators are selected at random to form a new committee, what is the probability that they will all be Republicans? What is the probability that they will all be Democrats? What is the probability that there will be 1 from each party, including the Independent?

   Source: New York Times Almanac
5. **Job Applications** Six men and seven women apply for two identical jobs. If the jobs are filled at random, find the following:
   a. The probability that both are filled by men.
   b. The probability that both are filled by women.
   c. The probability that one man and one woman are hired.
   d. The probability that the one man and one woman who are twins are hired.
6. **Defective Resistors** A package contains 12 resistors, 3 of which are defective. If 4 are selected, find the probability of getting
   a. 0 defective resistors
   b. 1 defective resistor
   c. 3 defective resistors
7. **Winning Tickets** At a meeting of 10 executives (7 women and 3 men), two door prizes are awarded. Find the probability that both prizes are won by men.
8. **Getting a Full House** Find the probability of getting a full house (3 cards of one denomination and 2 of another) when 5 cards are dealt from an ordinary deck.
9. **World-Class Orchestras** About.com's list of 20 World Class Orchestras includes the following from the United States: Boston Symphony Orchestra, Chicago Symphony Orchestra, Cleveland Orchestra, Los Angeles Philharmonic, New York Philharmonic, the Metropolitan Opera Orchestra, and the San Francisco Symphony. Choose 5 at random from the list of 20 for a benefit CD. What is the probability that the collection will include at least one group from the United States? At least 2 from the United States? That all 5 will be from the United States?
10. **Selecting Cards** The red face cards and the black cards numbered 2–9 are put into a bag. Four cards are drawn at random without replacement. Find the following probabilities.
    a. All 4 cards are red.
    b. 2 cards are red and 2 cards are black.
    c. At least 1 of the cards is red.
    d. All 4 cards are black.
11. **Socks in a Drawer** A drawer contains 11 identical red socks and 8 identical black socks. Suppose that you choose 2 socks at random in the dark.
    a. What is the probability that you get a pair of red socks?
    b. What is the probability that you get a pair of black socks?
    c. What is the probability that you get 2 unmatched socks?
    d. Where did the other red sock go?
12. **Selecting Books** Find the probability of selecting 3 science books and 4 math books from 8 science books and 9 math books. The books are selected at random.

13. **Rolling the Dice** If three dice are rolled, find the probability of getting a sum of 6.

14. **Football Team Selection** A football team consists of 20 freshmen and 20 sophomores, 15 juniors, and 10 seniors. Four players are selected at random to serve as captains. Find the probability that
    a. All 4 are seniors
    b. There is 1 each: freshman, sophomore, junior, and senior
    c. There are 2 sophomores and 2 freshmen
    d. At least 1 of the students is a senior

15. **Arrangement of Washers** Find the probability that if 5 different-sized washers are arranged in a row, they will be arranged in order of size.

16. **Poker Hands** Using the information in Exercise 70 in Section 4–4, find the probability of each poker hand.
    a. Royal flush
    b. Straight flush
    c. Four of a kind

17. **Plant Selection** All holly plants are dioecious—a male plant must be planted within 30 to 40 feet of the female plants in order to yield berries. A home improvement store has 12 unmarked holly plants for sale, 8 of which are female. If a homeowner buys 3 plants at random, what is the probability that berries will be produced?

Summary

In this chapter, the basic concepts of probability are explained.

- There are three basic types of probability: classical probability, empirical probability, and subjective probability. Classical probability uses sample spaces. Empirical probability uses frequency distributions, and subjective probability uses an educated guess to determine the probability of an event. The probability of any event is a number from 0 to 1. If an event cannot occur, the probability is 0. If an event is certain, the probability is 1. The sum of the probability of all the events in the sample space is 1. To find the probability of the complement of an event, subtract the probability of the event from 1. (4–1)
- Two events are mutually exclusive if they cannot occur at the same time; otherwise, the events are not mutually exclusive. To find the probability of two mutually exclusive events occurring, add the probability of each event. To find the probability of two events when they are not mutually exclusive, add the possibilities of the individual events and then subtract the probability that both events occur at the same time. These types of probability problems can be solved by using the addition rules. (4–2)
- Two events are independent if the occurrence of the first event does not change the probability of the second event occurring. Otherwise, the events are dependent. To find the probability of two independent events occurring, multiply the probabilities of each event. To find the probability that two dependent events occur, multiply the probability that the first event occurs by the probability that the second event occurs, given that the first event has already occurred. The complement of an event is found by selecting the outcomes in the sample space that are not involved in the outcomes of the event. These types of problems can be solved by using the multiplication rules and the complementary event rules. (4–3)
- Finally, when a large number of events can occur, the fundamental counting rule, the permutation rules, and the combination rule can be used to determine the number of ways that these events can occur. (4–4)
- The counting rules and the probability rules can be used to solve more-complex probability problems. (4–5)

Important Terms

classical probability 189
combination 232
complement of an event 192
compound event 189
conditional probability 215
dependent events 215
disjoint events 202
empirical probability 194
equally likely events 189
event 188
fundamental counting rule 226
independent events 213
law of large numbers 196
mutually exclusive events 202
outcome 186
permutation 229
probability 186
probability experiment 186
sample space 186
simple event 189
subjective probability 196
tree diagram 188
Venn diagrams 193

Important Formulas

Formula for classical probability:

$$P(E) = \frac{\text{number of outcomes in } E}{\text{total number of outcomes in sample space}} = \frac{n(E)}{n(S)}$$

Formula for empirical probability:

$$P(E) = \frac{\text{frequency for class}}{\text{total frequencies in distribution}} = \frac{f}{n}$$

Addition rule 1, for two mutually exclusive events:

$$P(A \text{ or } B) = P(A) + P(B)$$

Addition rule 2, for events that are not mutually exclusive:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Multiplication rule 1, for independent events:

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

Multiplication rule 2, for dependent events:

$$P(A \text{ and } B) = P(A) \cdot P(B|A)$$

Formula for conditional probability:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Formula for complementary events:

$$P(\overline{E}) = 1 - P(E) \quad \text{or} \quad P(E) = 1 - P(\overline{E}) \quad \text{or} \quad P(E) + P(\overline{E}) = 1$$

Fundamental counting rule: In a sequence of n events in which the first one has k_1 possibilities, the second event has k_2 possibilities, the third has k_3 possibilities, etc., the total number of possibilities of the sequence will be

$$k_1 \cdot k_2 \cdot k_3 \cdots k_n$$

Permutation rule 1: The number of permutations of n objects taking r objects at a time when order is important is

$${}_nP_r = \frac{n!}{(n-r)!}$$

Permutation rule 2: The number of permutations of n objects when r_1 objects are identical, r_2 objects are identical, . . . , r_p objects are identical is

$$\frac{n!}{r_1!\,r_1!\cdots r_p!}$$

Combination rule: The number of combinations of r objects selected from n objects when order is not important is

$${}_nC_r = \frac{n!}{(n-r)!r!}$$

Review Exercises

Section 4–1

1. **Rolling a Die** An eight-sided die is rolled. Find the probability of each.
   a. Getting a 6
   b. Getting a number larger than 5
   c. Getting an odd number
2. **Selecting a Card** When a card is selected from a deck, find the probability of getting
   a. A club
   b. A face card or a heart
   c. A 6 and a spade
   d. A king
   e. A red card
3. **Software Selection** The top-10 selling computer software titles in a recent year consisted of 3 for doing taxes, 5 antivirus or security programs, and 2 "other." Choose one title at random.
   a. What is the probability that it is not used for doing taxes?
   b. What is the probability that it is used for taxes or is one of the "other" programs?

   Source: www.nielsen.com
4. **Motor Vehicle Producers** The top five motor vehicle producers in the world are listed below with the number of vehicles produced in 2010 (in thousands of vehicles).

China	16,144
Japan	9,197
United States	7,632
Germany	5,700
South Korea	4,184

   Choose one vehicle at random;
   a. What is the probability that it was produced in the United States?
   b. What is the probability that it was not produced in Asia?
   c. What is the probability that it was produced in Germany or Japan?

   Source: World Almanac 2012.
5. **Exercise Preference** In a local fitness club, 32% of the members lifted weights to exercise, 41% preferred aerobics, and 6% preferred both. If a member is selected at random, find the probability that the member prefers another method (neither weights nor aerobics) to exercise.

6. **Rolling Two Dice** When two dice are rolled, find the probability of getting
   a. A sum of 5 or 6
   b. A sum greater than 9
   c. A sum less than 4 or greater than 9
   d. A sum that is divisible by 4
   e. A sum of 14
   f. A sum less than 13

Section 4–2

7. **New Cars** The probability that a new automobile has a backup camera is 0.6. The probability that a new automobile has a GPS system is 0.4. The probability that a new automobile has both a backup camera and a GPS is 0.2. If a new automobile is selected at random, find the probability that it has neither a backup camera nor a GPS.

8. **Breakfast Drink** In a recent survey, 18 people preferred milk, 29 people preferred coffee, and 13 people preferred juice as their primary drink for breakfast. If a person is selected at random, find the probability that the person preferred juice as her or his primary drink.

9. **Lawnmower and Weed Wacker Ownership** The probability that a homeowner owns a lawnmower is 0.7. The probability that a homeowner owns a weed wacker is 0.5. The probability that a home owner owns both a lawnmower and a weed wacker is 0.3. If a homeowner is selected at random, find the probability that he or she owns either a lawnmower or a weed wacker.

10. **Casino Gambling** The probability that a gambler plays table games is 0.32. The probability that a person plays the slot machines is 0.85. The probability that a person plays both is 0.15. A gambler can play more than one type of game. Find the probability that a gambler plays table games given that this person plays the slot machines.

11. **Online Course Selection** Roughly 1 in 6 students enrolled in higher education took at least one online course last fall. Choose 5 enrolled students at random. Find the probability that
    a. All 5 took online courses
    b. None of the 5 took a course online
    c. At least 1 took an online course

    Source: www.encarta.msn.com

12. **Purchasing Sweaters** During a sale at a men's store, 16 white sweaters, 3 red sweaters, 9 blue sweaters, and 7 yellow sweaters were purchased. If a customer is selected at random, find the probability that he bought
    a. A blue sweater
    b. A yellow or a white sweater
    c. A red, a blue, or a yellow sweater
    d. A sweater that was not white

Section 4–3

13. **Drawing Cards** Three cards are drawn from an ordinary deck *without* replacement. Find the probability of getting
    a. All black cards
    b. All spades
    c. All queens

14. **Coin Toss and Card Drawn** A coin is tossed and a card is drawn from a deck. Find the probability of getting
    a. A head and a 6
    b. A tail and a red card
    c. A head and a club

15. **Movie Releases** The top five countries for movie releases for a specific year are the United States with 471 releases, United Kingdom with 386, Japan with 79, Germany with 316, and France with 132. Choose 1 new release at random. Find the probability that it is
    a. European
    b. From the United States
    c. German or French
    d. German given that it is European

    Source: www.showbizdata.com

16. **Factory Output** A manufacturing company has three factories: X, Y, and Z. The daily output of each is shown here.

Product	Factory X	Factory Y	Factory Z
TVs	18	32	15
Stereos	6	20	13

    If 1 item is selected at random, find these probabilities.
    a. It was manufactured at factory X or is a stereo.
    b. It was manufactured at factory Y or factory Z.
    c. It is a TV or was manufactured at factory Z.

17. **Effectiveness of Vaccine** A vaccine has a 90% probability of being effective in preventing a certain disease. The probability of getting the disease if a person is not vaccinated is 50%. In a certain geographic region, 25% of the people get vaccinated. If a person is selected at random, find the probability that he or she will contract the disease.

18. **T-shirt Factories** Two T-shirt printing factories produce T-shirts for a local sports team. Factory A produces 60% of the shirts and factory B produces 40%. Five percent of the shirts from factory A are defective, and 6% of the shirts from factory B are defective. Choose 1 shirt at random. Given that the shirt is defective, what is the probability that it came from factory A?

19. **Car Purchase** The probability that Sue will live on campus and buy a new car is 0.37. If the probability that she will live on campus is 0.73, find the probability that she will buy a new car, given that she lives on campus.

20. **Applying Shipping Labels** Four unmarked packages have lost their shipping labels, and you must reapply them. What is the probability that you apply the labels and get all 4 of them correct? Exactly 3 correct? Exactly 2? At least 1 correct?

21. **Health Club Membership** Of the members of the Blue River Health Club, 43% have a lifetime membership and

exercise regularly (three or more times a week). If 75% of the club members exercise regularly, find the probability that a randomly selected member is a life member, given that he or she exercises regularly.

22. **Bad Weather** The probability that it snows and the bus arrives late is 0.023. José hears the weather forecast, and there is a 40% chance of snow tomorrow. Find the probability that the bus will be late, given that it snows.

23. **Education Level and Smoking** At a large factory, the employees were surveyed and classified according to their level of education and whether they smoked. The data are shown in the table.

	Educational level		
Smoking habit	Not high school graduate	High school graduate	College graduate
Smoke	6	14	19
Do not smoke	18	7	25

If an employee is selected at random, find these probabilities

a. The employee smokes, given that he or she graduated from college.

b. Given that the employee did not graduate from high school, he or she is a smoker.

24. **War Veterans** Approximately 11% of the civilian population are veterans. Choose 5 civilians at random. What is the probability that none are veterans? What is the probability that at least 1 is a veteran?

Source: www.factfinder.census.gov

25. **Television Sets** If 98% of households have at least one television set and 4 households are selected, find the probability that at least one household has a television set.

26. **Chronic Sinusitis** The U.S. Department of Health and Human Services reports that 15% of Americans have chronic sinusitis. If 5 people are selected at random, find the probability that at least 1 has chronic sinusitis.

Source: 100% American

Section 4–4

27. **Motorcycle License Plates** If a motorcycle license plate consists of two letters followed by three digits, how many different license plates can be made if repetitions are allowed? How many different license plates can be made if repetitions are not allowed? How many license plates can be made if repetitions are allowed in the digits but not in the letters?

28. **Types of Copy Paper** White copy paper is offered in 5 different strengths and 11 different degrees of brightness, recycled or not, and acid-free or not. How many different types of paper are available for order?

29. **Baseball Players** How many ways can 3 outfielders and 4 infielders be chosen from 5 outfielders and 7 infielders?

30. **Carry-on Items** The following items are allowed as airline carry-on items: (1) safety razors, (2) eyedrops and saline, (3) nail clippers and tweezers, (4) blunt-tipped scissors, (5) mobile phones, (6) umbrellas, (7) common lighters, (8) beverages purchased after security screening, and (9) musical instruments. Suppose that your airline allows only 6 of these items. In how many ways can you pick 3 not to take?

31. **Names for Boys** The top 10 names for boys in America in 2005 were Ethan, Jacob, Ryan, Matthew, Tyler, Jack, Joshua, Andrew, Noah, and Michael. The top 10 names for 2015 are Liam, Noah, Ethan, Mason, Lucas, Logan, Oliver, Jackson, Aiden, and Jacob. In how many ways can you choose 5 names from these lists?

32. **Committee Representation** There are 6 Republican, 5 Democrat, and 4 Independent candidates. How many different ways can a committee of 3 Republicans, 2 Democrats, and 1 Independent be selected?

33. **Song Selections** A promotional MP3 player is available with the capacity to store 100 songs, which can be reordered at the push of a button. How many different arrangements of these songs are possible? (*Note:* Factorials get very big, very fast! How large a factorial will your calculator calculate?)

34. **Employee Health Care Plans** A new employee has a choice of 5 health care plans, 3 retirement plans, and 2 different expense accounts. If a person selects 1 of each option, how many different options does she or he have?

35. **Course Enrollment** There are 12 students who wish to enroll in a particular course. There are only 4 seats left in the classroom. How many different ways can 4 students be selected to attend the class?

36. **Candy Selection** A candy store allows customers to select 3 different candies to be packaged and mailed. If there are 13 varieties available, how many possible selections can be made?

37. **House Numbers** A home improvement store has the following house numbers: 335666 left after a sale. How many different six-digit house numbers can be made from those numbers?

38. **Word Permutations** From which word can you make more permutations, *MATHEMATICS* or *PROBABILITY*? How many more?

39. **Movie Selections** If a person can select one movie for each week night except Saturday, how many different ways can the selection be made if the person can select from 16 movies?

40. **Course Selection** If a student can select one of 3 language courses, one of 5 mathematics courses,

and one of 4 history courses, how many different schedules can be made?

Section 4–5

41. **Catalog ID Numbers** In a catalog, movies are identified by an ID number—2 letters followed by 3 digits. Repetitions are permitted. How many catalog numbers can be made? What is the probability that the ID number is divisible by 5? (Include the 2 letters in the ID number.)

42. **License Plates** A certain state's license plate has 3 letters followed by 4 numbers. Repeats are not allowed for the letters, but they are for the numbers. How many such license plates are possible? If they are issued at random, what is the probability that the 3 letters are 3 consecutive letters in alphabetical order?

43. **Territorial Selection** Several territories and colonies today are still under the jurisdiction of another country. France holds the most with 16 territories, the United Kingdom has 15, the United States has 14, and several other countries have territories as well. Choose 3 territories at random from those held by France, the United Kingdom, and the United States. What is the probability that all 3 belong to the same country?

Source: www.infoplease.com

44. **Yahtzee** Yahtzee is a game played with 5 dice. Players attempt to score points by rolling various combinations. When all 5 dice show the same number, it is called a *Yahtzee* and scores 50 points for the first one and 100 points for each subsequent Yahtzee in the same game. What is the probability that a person throws a Yahtzee on the very first roll? What is the probability that a person throws two Yahtzees on two successive turns?

45. **Personnel Classification** For a survey, a subject can be classified as follows:

Gender: male or female

Marital status: single, married, widowed, divorced

Occupation: administration, faculty, staff

Draw a tree diagram for the different ways a person can be classified.

STATISTICS TODAY

Would You Bet Your Life?—Revisited

In his book *Probabilities in Everyday Life*, John D. McGervey states that the chance of being killed on any given commercial airline flight is almost 1 in 1 million and that the chance of being killed during a transcontinental auto trip is about 1 in 8000. The corresponding probabilities are 1/1,000,000 = 0.000001 as compared to 1/8000 = 0.000125. Since the second number is 125 times greater than the first number, you have a much higher risk driving than flying across the United States.

Chapter Quiz

Determine whether each statement is true or false. If the statement is false, explain why.

1. Subjective probability has little use in the real world.
2. Classical probability uses a frequency distribution to compute probabilities.
3. In classical probability, all outcomes in the sample space are equally likely.
4. When two events are not mutually exclusive, $P(A \text{ or } B) = P(A) + P(B)$.
5. If two events are dependent, they must have the same probability of occurring.
6. An event and its complement can occur at the same time.
7. The arrangement ABC is the same as BAC for combinations.
8. When objects are arranged in a specific order, the arrangement is called a combination.

Select the best answer.

9. The probability that an event happens is 0.42. What is the probability that the event won't happen?

   a. −0.42 *c.* 0

   b. 0.58 *d.* 1

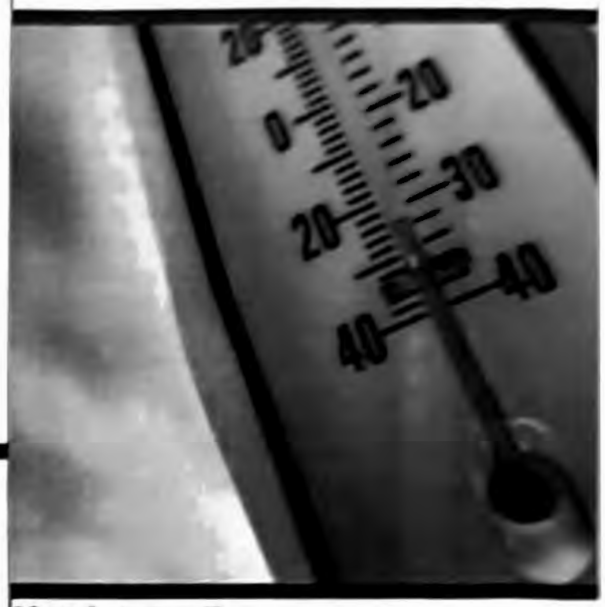
10. When a meteorologist says that there is a 30% chance of showers, what type of probability is the person using?

    a. Classical *c.* Relative

    b. Empirical *d.* Subjective

5

Correlation and Regression



© Design Pics Inc./Alamy RF

STATISTICS TODAY

Can Temperature Predict Homicides?

Over the last years, researchers have been interested in the relationship between increasing temperatures and increasing homicide rates. To test this relationship, the author selected a city in the Midwest and obtained the average monthly temperatures for that city as well as the number of homicides committed each month for the year 2015. The data are shown.

Month	Average High Temperature	Homicides
January	32	32
February	38	20
March	47	35
April	59	35
May	70	49
June	80	49
July	84	53
August	83	56
September	76	62
October	64	29
November	49	36
December	37	32

Source: Chicago Tribune

Using the statistical methods described in this chapter, you will be able to answer these questions:

1. Is there a linear relationship between the monthly average temperatures and the number of homicides committed during the month?
2. If so, how strong is the relationship between the average monthly temperature and the number of homicides committed?
3. If a relationship exists, can it be said that an increase in temperatures will cause an increase in the number of homicides occurring in that city?

See Statistics Today—Revisited at the end of the chapter for the answers to these questions.

OUTLINE

Introduction

5–1 Scatter Plots and Correlation

5–2 Regression

5–3 Coefficient of Determination and Standard Error of the Estimate

5–4 Multiple Regression (Optional)

Summary

OBJECTIVES

After completing this chapter, you should be able to

1. Draw a scatter plot for a set of ordered pairs.
2. Compute the correlation coefficient.
3. Test the hypothesis H_0: $\rho = 0$.
4. Compute the equation of the regression line.
5. Compute the coefficient of determination.
6. Compute the standard error of the estimate.
7. Find a prediction interval.
8. Be familiar with the concept of multiple regression.

Introduction

In Chapters 7 and 8, two areas of inferential statistics—confidence intervals and hypothesis testing—were explained. Another area of inferential statistics involves determining whether a relationship exists between two or more numerical or quantitative variables. For example, a businessperson may want to know whether the volume of sales for a given month is related to the amount of advertising the firm does that month. Educators are interested in determining whether the number of hours a student studies is related to the student's score on a particular exam. Medical researchers are interested in questions such as, Is caffeine related to heart damage? or Is there a relationship between a person's age and his or her blood pressure? A zoologist may want to know whether the birth weight of a certain animal is related to its life span. These are only some of the many questions that can be answered by using the techniques of correlation and regression analysis.

Unusual Stat

A person walks on average 100,000 miles in his or her lifetime. This is about 3.4 miles per day.

The purpose of this chapter then is to answer these questions statistically:

1. Are two or more variables linearly related?
2. If so, what is the strength of the relationship?
3. What type of relationship exists?
4. What kind of predictions can be made from the relationship?

5–1 Scatter Plots and Correlation

OBJECTIVE 1

Draw a scatter plot for a set of ordered pairs.

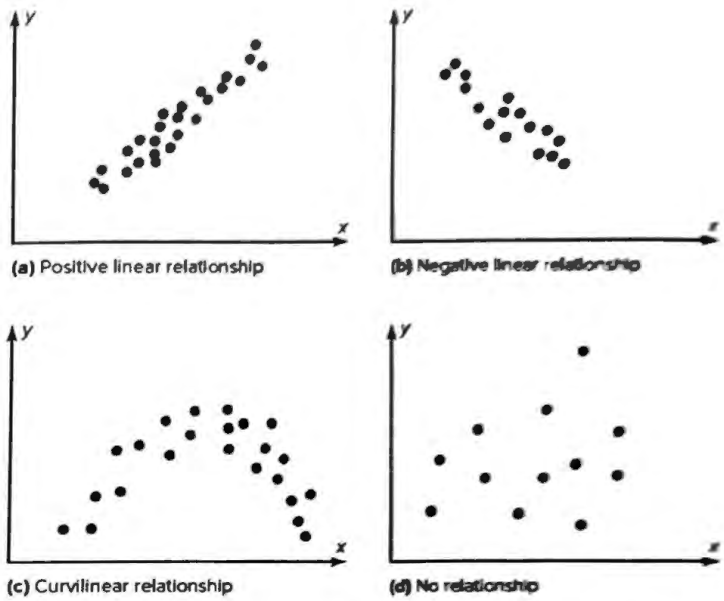
In simple correlation and regression studies, the researcher collects data on two numerical or quantitative variables to see whether a relationship exists between the variables. For example, if a researcher wishes to see whether there is a relationship between number of hours of study and test scores on an exam, she must select a random sample of students, determine the number of hours each studied, and obtain their grades on the exam. A table can be made for the data, as shown here.

Student	Hours of study x	Grade y (%)
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

The two variables for this study are called the **independent variable** and the **dependent variable.** The independent variable is the variable in regression that can be controlled or manipulated. In this case, the number of hours of study is the independent variable and is designated as the x variable. The dependent variable is the variable in regression that cannot be controlled or manipulated. The grade the student received on the exam is the dependent variable, designated as the y variable. The reason for this distinction between the variables is that you assume that the grade the student earns *depends* on the number of hours the student studied. Also, you assume that, to some extent, the student can regulate or *control* the number of hours he or she studies for the exam. The independent variable is also known as the *explanatory variable,* and the dependent variable is also called the *response variable.*

The determination of the x and y variables is not always clear-cut and is sometimes an arbitrary decision. For example, if a researcher studies the effects of age on a person's blood pressure, the researcher can generally assume that age affects blood pressure. Hence, the variable *age* can be called the *independent variable,* and the variable *blood pressure* can be called the *dependent variable.* On the other hand, if a researcher is studying the attitudes of husbands on a certain issue and the attitudes of their wives on the same issue, it is difficult to say which variable is the independent variable and which is the dependent variable. In this study, the researcher can arbitrarily designate the variables as independent and dependent.

FIGURE 5–1
Types of Relationships



The independent and dependent variables can be plotted on a graph called a *scatter plot.* The independent variable x is plotted on the horizontal axis, and the dependent variable y is plotted on the vertical axis.

A scatter plot is a graph of the ordered pairs (x, y) of numbers consisting of the independent variable x and the dependent variable y.

The scatter plot is a visual way to describe the nature of the relationship between the independent and dependent variables. The scales of the variables can be different, and the coordinates of the axes are determined by the smallest and largest data values of the variables.

Researchers look for various types of patterns in scatter plots. For example, in Figure 5–1(a), the pattern in the points of the scatter plot shows a **positive linear relationship.** Here, as the values of the independent variable (x variable) increase, the values of the dependent variable (y variable) increase. Also, the points form somewhat of a straight line going in an upward direction from left to right.

The pattern of the points of the scatter plot shown in Figure 5–1(b) shows a **negative linear relationship.** In this case, as the values of the independent variable increase, the values of the dependent variable decrease. Also, the points show a somewhat straight line going in a downward direction from left to right.

The pattern of the points of the scatter plot shown in Figure 5–1(c) shows some type of a nonlinear relationship or a curvilinear relationship.

Finally, the scatter plot shown in Figure 5–1(d) shows basically no relationship between the independent variable and the dependent variable since no pattern (line or curve) can be seen.

The procedure table for drawing a scatter plot is given next.

Procedure Table

Drawing a Scatter Plot

Step 1	Draw and label the x and y axes.
Step 2	Plot each point on the graph.
Step 3	Determine the type of relationship (if any) that exists for the variables.

The procedure for drawing a scatter plot is shown in Examples 10–1 through 10–3.

EXAMPLE 5–1 Car Rental Companies Construct a scatter plot for the data shown for car rental companies in the United States for a recent year.

Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

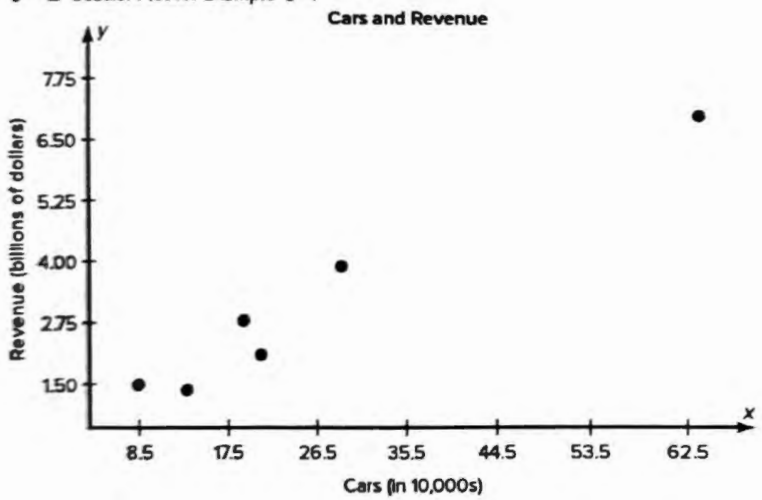
Source: Auto Rental News.

SOLUTION

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph, as shown in Figure 5–2.

FIGURE 5–2 Scatter Plot for Example 5–1



Step 3 Determine the type of relationship (if any) that exists.

In this example, it looks as if a positive linear relationship exists between the number of cars that an agency owns and the total revenue that is made by the company.

EXAMPLE 5–2 Absences and Final Grades

Construct a scatter plot for the data obtained in a study on the number of absences and the final grades of seven randomly selected students from a statistics class. The data are shown here.

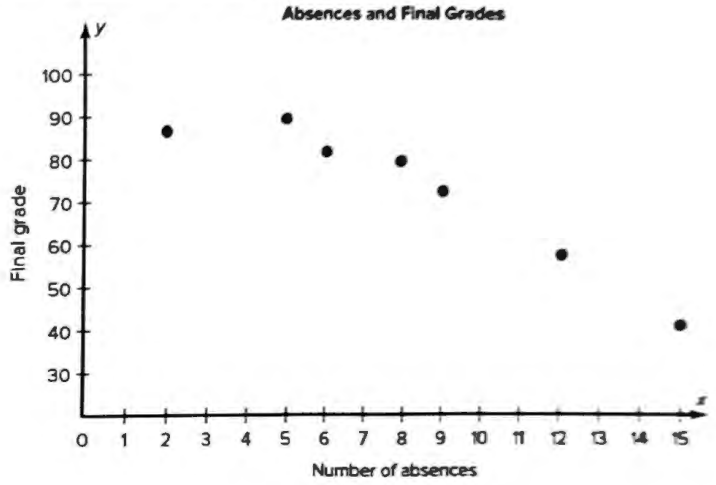
Student	Number of absences x	Final grade y (%)
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78

SOLUTION

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph, as shown in Figure 10–3.

FIGURE 5 – 3 Scatter Plot for Example 5–2



Step 3 Determine the type of relationship (if any) that exists.

In this example, it looks as if a negative linear relationship exists between the number of student absences and the final grade of the students.

EXAMPLE 5–3 Number of Teachers and Pupils per Teacher

A researcher wishes to see if there is a relationship between the number of pupils per teacher and the number of teachers (in thousands) employed by the school district. She randomly selects 10 school districts throughout the United States. The data are shown.

School district	Number of teachers (in thousands)	Pupils per teacher
1	7	12.4
2	34	14.3
3	9	14.3
4	8	9.2
5	16	18.3
6	15	12.1
7	6	12.3
8	14	12.4
9	32	15.2
10	10	13.4

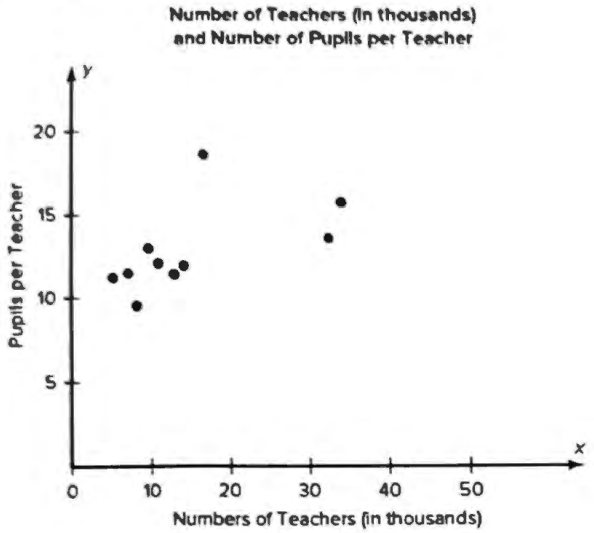
Source: U.S. Department of Education.

SOLUTION

Step 1 Draw and label the x and y axes.

Step 2 Plot each point on the graph as shown in Figure 10–4.

FIGURE 5–4 Scatter Plot for Example 10–3



Step 3 Determine the type of relationship (if any) that exists.

In this case, there is no indication of a strong positive or negative linear relationship between the number of pupils per teacher and the number of teachers (in thousands) in a school district.

Correlation

OBJECTIVE 2

Compute the correlation coefficient.

Correlation Coefficient Statisticians use a measure called the *correlation coefficient* to determine the strength of the linear relationship between two variables. There are several types of correlation coefficients.

The population correlation coefficient denoted by the Greek letter ρ is the correlation computed by using all possible pairs of data values (x, y) taken from a population.

The linear correlation coefficient computed from the sample data measures the strength and direction of a linear relationship between two quantitative variables. The symbol for the sample correlation coefficient is r.

The linear correlation coefficient explained in this section is called the **Pearson product moment correlation coefficient (PPMC),** named after statistician Karl Pearson, who pioneered the research in this area.

The *range of the linear correlation coefficient* is from -1 to $+1$. If there is a *strong positive linear relationship* between the variables, the value of r will be close to $+1$. If there is a *strong negative linear relationship* between the variables, the value of r will be close to -1. When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0. See Figure 10–5. When the value of r is 0 or close to zero, it implies only that there is no linear relationship between the variables. The data may be related in some other nonlinear way.

FIGURE 5–5 Range of Values for the Correlation Coefficient

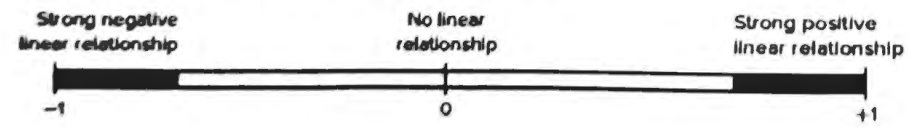
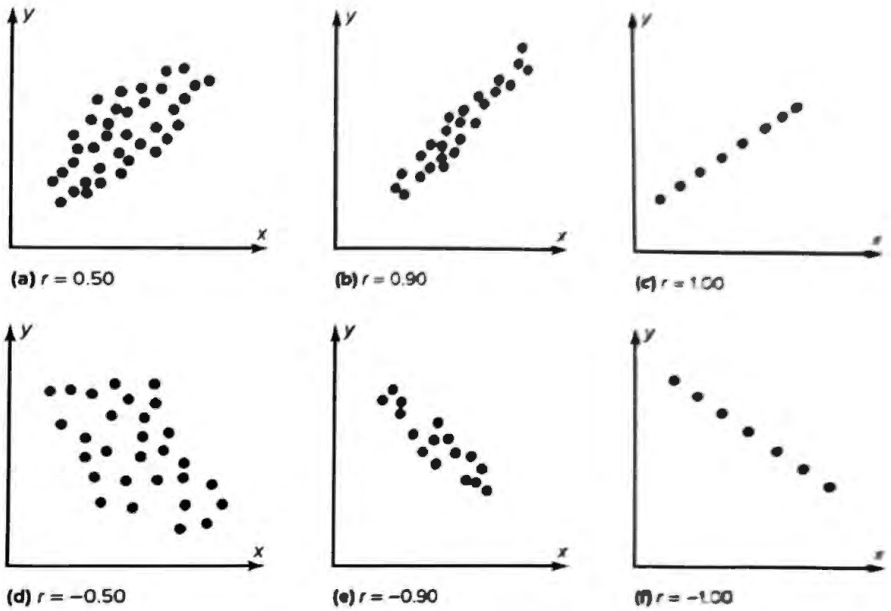


FIGURE 5–6
Relationship Between the Correlation Coefficient and the Scatter Plot



Properties of the Linear Correlation Coefficient

1. The correlation coefficient is a unitless measure.
2. The value of r will always be between −1 and +1 inclusively. That is, $-1 \leq r \leq 1$.
3. If the values of x and y are interchanged, the value of r will be unchanged.
4. If the values of x and/or y are converted to a different scale, the value of r will be unchanged.
5. The value of r is sensitive to outliers and can change dramatically if they are present in the data.

The graphs in Figure 10–6 show the relationship between the correlation coefficients and their corresponding scatter plots. Notice that as the value of the correlation coefficient increases from 0 to +1 (parts *a*, *b*, and *c*), data values become closer to a straight line and to an increasingly strong relationship. As the value of the correlation coefficient decreases from 0 to −1 (parts *d*, *e*, and *f*), the data values also become closer to a straight line. Again this suggests a stronger relationship.

Assumptions for the Correlation Coefficient

1. The sample is a random sample.
2. The data pairs fall approximately on a straight line and are measured at the interval or ratio level.
3. The variables have a bivariate normal distribution. (This means that given any specific value of x, the y values are normally distributed; and given any specific value of y, the x values are normally distributed.)

In this book, the assumptions will be stated in the exercises; however, when encountering statistics in other situations, you must check to see that these assumptions have been met before proceeding.

There are several ways to compute the value of the correlation coefficient. One method is to use the formula shown here.

Formula for the Linear Correlation Coefficient r

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

where n is the number of data pairs.

Rounding Rule for the Correlation Coefficient Round the value of r to three decimal places.

The formula looks somewhat complicated, but using a table to compute the values, as shown in Example 10–4, makes it somewhat easier to determine the value of r.

There are no units associated with r, and the value of r will remain unchanged if the x and y values are switched.

The procedure for finding the value of the linear correlation coefficient is given next.

Procedure Table

Finding the Value of the Linear Correlation Coefficient

Step 1 Make a table as shown.

x	y	xy	x^2	y^2

Step 2 Place the values of x in the x column and the values of y in the y column.
Multiply each x value by the corresponding y value, and place the products in the xy column.
Square each x value and place the squares in the x^2 column.
Square each y value and place the squares in the y^2 column.
Find the sum of each column.

Step 3 Substitute in the formula and find the value for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

where n is the number of data pairs.

EXAMPLE 5–4 Car Rental Companies

Compute the linear correlation coefficient for the data in Example 5–1.

SOLUTION

Step 1 Make a table as shown here.

Company	Cars x (in ten thousands)	Revenue y (in billions)	xy	x^2	y^2
A	63.0	$7.0			
B	29.0	3.9			
C	20.8	2.1			
D	19.1	2.8			
E	13.4	1.4			
F	8.5	1.5			

Step 2 Find the values of xy, x^2, and y^2, and place these values in the corresponding columns of the table.

The completed table is shown.

Company	Cars x (in 10,000s)	Revenue y (in billions of dollars)	xy	x^2	y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	12.75	72.25	2.25
	$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

Step 3 Substitute in the formula and solve for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

The linear correlation coefficient suggests a strong positive linear relationship between the number of cars a rental agency has and its annual revenue. That is, the more cars a rental agency has, the more annual revenue the company will have.

EXAMPLE 5–5 Absences and Final Grades

Compute the value of the linear correlation coefficient for the data obtained in the study of the number of absences and the final grade of the seven students in the statistics class given in Example 5–2.

SOLUTION

Step 1 Make a table.

Step 2 Find the values of xy, x^2, and y^2; place these values in the corresponding columns of the table.

Student	Number of absences x	Final grade y (%)	xy	x^2	y^2
A	6	82	492	36	6,724
B	2	86	172	4	7,396
C	15	43	645	225	1,849
D	9	74	666	81	5,476
E	12	58	696	144	3,364
F	5	90	450	25	8,100
G	8	78	624	64	6,084
	$\Sigma x = 57$	$\Sigma y = 511$	$\Sigma xy = 3745$	$\Sigma x^2 = 579$	$\Sigma y^2 = 38,993$

Step 3 Substitute in the formula and solve for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(7)(3745) - (57)(511)}{\sqrt{[(7)(579) - (57)^2][(7)(38,993) - (511)^2]}} = -0.944$$

The value of r suggests a strong negative linear relationship between a student's final grade and the number of absences a student has. That is, the more absences a student has, the lower is his or her grade.

EXAMPLE 5–6 Numbers of Teachers and Pupils per Teacher

Compute the value of the linear correlation coefficient for the data given in Example 5–3 for the number of teachers (in thousands) and the number of pupils per teacher.

SOLUTION

Step 1 Make a table.

Step 2 Find the values of xy, x^2, and y^2; place these values in the corresponding columns of the tables.

School district	Number of teachers, x	Pupils per teacher, y	xy	x^2	y^2
1	7	12.4	86.8	49	153.76
2	34	14.3	486.2	1156	204.49
3	9	14.3	128.7	81	204.49
4	8	9.2	73.6	64	84.64
5	16	18.3	292.8	256	334.89
6	15	12.1	181.5	225	146.41
7	6	12.3	73.8	36	151.29
8	14	12.4	173.6	196	153.76
9	32	15.2	486.4	1024	231.04
10	10	13.4	134	100	179.56
	$\Sigma x = 151$	$\Sigma y = 133.9$	$\Sigma xy = 2117.4$	$\Sigma x^2 = 3187$	$\Sigma y^2 = 1844.33$

Step 3 Substitute in the formula and solve for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{10(2117.4) - (151)(133.9^2)}{\sqrt{[10(3187) - (151)^2][10(1844.33) - (133.9^2)]}}$$

$$= \frac{955.1}{\sqrt{(9069)(514.09)}}$$

$$= \frac{955.1}{2159.23} = 0.442$$

The value of r indicates a weak positive linear relationship between the number of teachers (in thousands) employed and the number of pupils per teacher.

In Example 5–4, the value of r was high (close to 1.00); in Example 5–6, the value of r was much lower (close to 0). This question then arises, When is the value of r due to chance, and when does it suggest a significant linear relationship between the variables? This question will be answered next.

OBJECTIVE 3

Test the hypothesis H_0: $\rho = 0$

The Significance of the Correlation Coefficient As stated before, the range of the correlation coefficient is between −1 and + 1. When the value of r is near + 1 or −1, there is a strong linear relationship. When the value of r is near 0, the linear relationship is weak or nonexistent. Since the value of r is computed from data obtained from samples, there are two possibilities when r is not equal to zero: either the value of r is high enough to conclude that there is a significant linear relationship between the variables, or the value of r is due to chance.

To make this decision, you use a hypothesis-testing procedure. The traditional method is similar to the one used in previous chapters.

Step 1 State the hypotheses.

Step 2 Find the critical values.

Step 3 Compute the test value.

Step 4 Make the decision.

Step 5 Summarize the results.

The sample correlation coefficient can then be used as an estimator of ρ if the following assumptions are valid.

Assumptions for Testing the Significance of the Linear Correlation Coefficient

1. The data are quantitative and are obtained from a simple random sample.
2. The scatter plot shows that the data are approximately linearly related.
3. There are no outliers in the data.
4. The variables x and y must come from normally distributed populations.

In this book, the assumptions will be stated in the exercises; however, when encountering statistics in other situations, you must check to see that these assumptions have been met before proceeding.

In hypothesis testing, one of these is true:

H_0: $\rho = 0$ This null hypothesis means that there is no correlation between the x and y variables in the population.

H_1: $\rho \neq 0$ This alternative hypothesis means that there is a significant correlation between the variables in the population.

When the null hypothesis is rejected at a specific level, it means that there is a significant difference between the value of r and 0. When the null hypothesis is not rejected, it means that the value of r is not significantly different from 0 (zero) and is probably due to chance.

Several methods can be used to test the significance of the correlation coefficient. Three methods will be shown in this section. The first uses the t test.

Interesting Fact

Scientists think that a person is never more than 3 feet away from a spider at any given time!

Historical Notes

A mathematician named Karl Pearson (1857–1936) became interested in Francis Galton's work and saw that his correlation and regression theory could be applied to other areas besides heredity. Pearson developed the correlation coefficient that bears his name.

Formula for the t Test for the Correlation Coefficient

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

with degrees of freedom equal to $n - 2$, where n is the number of ordered pairs (x, y).

You do not have to identify the claim here, since the question will always be whether there is a significant linear relationship between the variables.

The two-tailed critical values are used. These values are found in Table F in Appendix A. Also, when you are testing the significance of a correlation coefficient, both variables x and y must come from normally distributed populations.

EXAMPLE 5–7

Test the significance of the correlation coefficient found in Example 5–4. Use $\alpha = 0.05$ and $r = 0.982$.

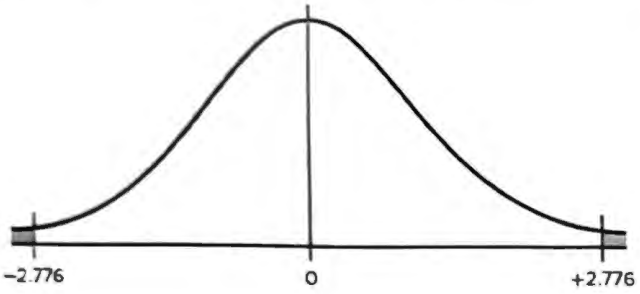
SOLUTION

Step 1 State the hypotheses.

$$H_0: \rho = 0 \quad \text{and} \quad H_1: \rho \neq 0$$

Step 2 Find the critical values. Since $\alpha = 0.05$ and there are $6 - 2 = 4$ degrees of freedom, the critical values obtained from Table F are ± 2.776, as shown in Figure 10–7.

FIGURE 5–7
Critical Values for Example 10–7

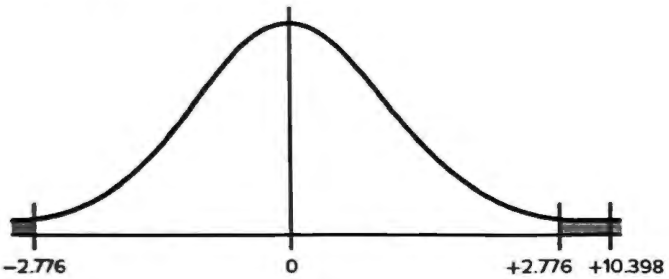


Step 3 Compute the test value.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.982\sqrt{\frac{6-2}{1-(0.982)^2}} = 10.398$$

Step 4 Make the decision. Reject the null hypothesis, since the test value falls in the critical region, as shown in Figure 10–8.

FIGURE 5–8
Test Value for Example 5–7



Step 5 Summarize the results. There is a significant relationship between the number of cars a rental agency owns and its annual income.

The second method that can be used to test the significance of r is the P-value method. The method is the same as that shown in Chapters 8 and 9. It uses the following steps.

Step 1 State the hypotheses.

Step 2 Find the test value. (In this case, use the t test.)

Step 3 Find the P-value. (In this case, use Table F.)

Step 4 Make the decision.

Step 5 Summarize the results.

Consider an example where $t = 4.059$, d.f. $= 4$, and $\alpha = 0.05$. Using Table F with d.f. $= 4$ and the row Two tails, the value 4.059 falls between 3.747 and 4.604; hence, $0.01 < P\text{-value} < 0.02$. (The P-value obtained from a calculator is 0.015.) That is, the P-value falls between 0.01 and 0.02. The decision, then, is to reject the null hypothesis since $P\text{-value} < 0.05$.

The third method of testing the significance of r is to use Table I in Appendix A. This table shows the values of the correlation coefficient that are significant for a specific α level and a specific number of degrees of freedom. For example, for 7 degrees of freedom and $\alpha = 0.05$, the table gives a critical value of 0.666. Any value of r greater than +0.666 or less than −0.666 will be significant, and the null hypothesis will be rejected. See Figure 10–9. When Table I is used, you need not compute the t test value. Table I is for two-tailed tests only.

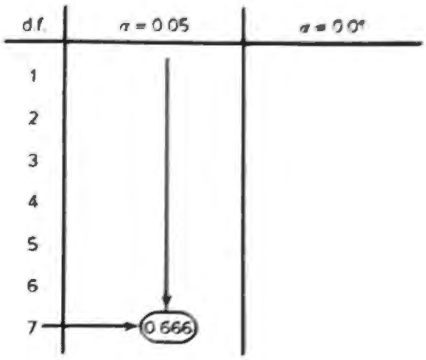


FIGURE 5–9
Finding the Critical Value from Table I

EXAMPLE 5–8

Using Table I, test the significance at $\alpha = 0.01$ of the correlation coefficient $r = 0.442$ obtained in Example 10–6.

SOLUTION

$$H_0: \rho = 0 \quad \text{and} \quad \rho \neq 0$$

Since the sample size is 10, there are $n - 2 = 10 - 2 = 8$ degrees of freedom. The critical values obtained from Table I at $\alpha = 0.01$ and 8 degrees of freedom are ± 0.765. Since $0.442 < 0.765$, the decision is not to reject the null hypothesis. See Figure 10–10. Hence, there is not enough evidence to say that there is a significant linear relationship.

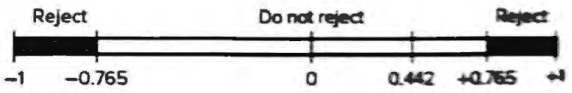


FIGURE 5–10
Rejection and Nonrejection Regions for Example 5–8

Generally, significance tests for correlation coefficients are two-tailed; however, they can be one-tailed. For example, if a researcher hypothesized a positive linear relationship between two variables, the hypotheses would be

$$H_0: \rho = 0$$
$$H_1: \rho > 0$$

If the researcher hypothesized a negative linear relationship between two variables, the hypotheses would be

$$H_0: \rho = 0$$
$$H_1: \rho < 0$$

In these cases, the t tests and the P-value tests would be one-tailed. Also, tables such as Table I are available for one-tailed tests. In this book, the examples and exercises will involve two-tailed tests.

Correlation and Causation Researchers must understand the nature of the linear relationship between the independent variable x and the dependent variable y. When a hypothesis test indicates that a significant linear relationship exists between the variables, researchers must consider the possibilities outlined next.

When two variables are highly correlated, item 3 in the box states that there exists a possibility that the correlation is due to a third variable. If this is the case and the third

Possible Relationships Between Variables

When the null hypothesis has been rejected for a specific α value, any of the following five possibilities can exist.

1. *There is a direct cause-and-effect relationship between the variables.* That is, x causes y. For example, water causes plants to grow, poison causes death, and heat causes ice to melt.
2. *There is a reverse cause-and-effect relationship between the variables.* That is, y causes x. For example, suppose a researcher believes excessive coffee consumption causes nervousness, but the researcher fails to consider that the reverse situation may occur. That is, it may be that an extremely nervous person craves coffee to calm his or her nerves.
3. *The relationship between the variables may be caused by a third variable.* For example, if a statistician correlated the number of deaths due to drowning and the number of cans of soft drink consumed daily during the summer, he or she would probably find a significant relationship. However, the soft drink is not necessarily responsible for the deaths, since both variables may be related to heat and humidity.
4. *There may be a complexity of interrelationships among many variables.* For example, a researcher may find a significant relationship between students' high school grades and college grades. But there probably are many other variables involved, such as IQ, hours of study, influence of parents, motivation, age, and instructors.
5. *The relationship may be coincidental.* For example, a researcher may be able to find a significant relationship between the increase in the number of people who are exercising and the increase in the number of people who are committing crimes. But common sense dictates that any relationship between these two values must be due to coincidence.

variable is unknown to the researcher or not accounted for in the study, it is called a **lurking variable.** An attempt should be made by the researcher to identify such variables and to use methods to control their influence.

It is important to restate the fact that even if the correlation between two variables is high, it does not necessarily mean causation. There are other possibilities, such as lurking variables or just a coincidental relationship.

Also, you should be cautious when the data for one or both of the variables involve averages rather than individual data. It is not wrong to use averages, but the results cannot be generalized to individuals since averaging tends to smooth out the variability among individual data values. The result could be a higher correlation than actually exists.

Thus, when the null hypothesis is rejected, the researcher must consider all possibilities and select the appropriate one as determined by the study. Remember, correlation does not necessarily imply causation.

Applying the Concepts 5–1

Stopping Distances

In a study on speed control, it was found that the main reasons for regulations were to make traffic flow more efficient and to minimize the risk of danger. An area that was focused on in the study was the distance required to completely stop a vehicle at various speeds. Use the following table to answer the questions.

MPH	Braking distance (feet)
20	20
30	45
40	81
50	133
60	205
80	411

Assume MPH is going to be used to predict stopping distance.

1. Which of the two variables is the independent variable?
2. Which is the dependent variable?
3. What type of variable is the independent variable?
4. What type of variable is the dependent variable?
5. Construct a scatter plot for the data.
6. Is there a linear relationship between the two variables?
7. Redraw the scatter plot, and change the distances between the independent-variable numbers. Does the relationship look different?
8. Is the relationship positive or negative?
9. Can braking distance be accurately predicted from MPH?
10. List some other variables that affect braking distance.
11. Compute the value of r.
12. Is r significant at $\alpha = 0.05$?

See page 604 for the answers.

Exercises 5–1

1. What is meant by the statement that two variables are related?

2. How is a linear relationship between two variables measured in statistics? Explain.

3. What is the symbol for the sample correlation coefficient? The population correlation coefficient?

4. What is the range of values for the correlation coefficient?

5. What is meant when the relationship between the two variables is called positive? Negative?

6. Give examples of two variables that are positively correlated and two that are negatively correlated.

7. What is the diagram of the independent and dependent variables called? Why is drawing this diagram important?

8. What is the name of the correlation coefficient used in this section?

9. What statistical test is used to test the significance of the correlation coefficient?

10. When two variables are correlated, can the researcher be sure that one variable causes the other? Why or why not?

For Exercises 11 through 27, perform the following steps.

a. Draw the scatter plot for the variables.
b. Compute the value of the correlation coefficient.
c. State the hypotheses.
d. Test the significance of the correlation coefficient at $\alpha = 0.05$, using Table I.
e. Give a brief explanation of the type of relationship. Assume all assumptions have been met.

11. **Crimes** The number of murders and robberies per 100,000 population for a random selection of states is shown. Is there a linear relationship between the variables?

Murders	2.4	2.7	5.6	2.6	2.1	3.3	6.6	5.7
Robberies	25.3	14.3	151.6	91.1	80	49	173	95.8

Source: Time Almanac.

12. **Oil and Gas Prices** The average gasoline price per gallon (in cities) and the cost of a barrel of oil are shown for a random selection of weeks in 2015. Is there a linear relationship between the variables?

Oil ($)	51.91	60.65	59.56	52.86	45.12	44.21
Gasoline ($)	1.97	1.96	2.06	2.04	2.00	1.99

(The information in this exercise will be used for Exercise 12 in Section 10–2.)

Source: Trading Economics.

13. **Commercial Movie Releases** The yearly data have been published showing the number of releases for each of the commercial movie studios and the gross receipts for those studios thus far. Based on these data, can it be concluded that there is a linear relationship between the number of releases and the gross receipts?

No. of releases x	361	270	306	22	35	10	9	12	21
Gross receipts y (million $)	3844	1962	1371	1064	934	244	188	156	125

(The information in this exercise will be used for Exercises 13 and 36 in Section 10–2 and Exercises 15 and 19 in Section 10–3.)

Source: www.showbizdata.com

14. Forest Fires and Acres Burned An environmentalist wants to determine the relationships between the numbers (in thousands) of forest fires over the year and the number (in hundred thousands) of acres burned. The data for 8 recent years are shown. Describe the relationship.

Number of fires x	72	69	58	47	84	62	57	45
Number of acres burned y	62	42	19	26	51	15	30	15

Source: National Interagency Fire Center

(The information in this exercise will be used for Exercise 14 in Section 10–2 and Exercises 16 and 20 in Section 10–3.)

15. Alumni Contributions The director of an alumni association for a small college wants to determine whether there is any type of relationship between the amount of an alumnus's contribution (in dollars) and the number of years the alumnus has been out of school. The data follow.

Years x	1	5	3	10	7	6
Contribution y	500	100	300	50	75	80

(The information is used for Exercises 15, 36, and 37 in Section 10–2 and Exercises 17 and 21 in Section 10–3.)

16. State Debt and Per Capita Tax An economics student wishes to see if there is a relationship between the amount of state debt per capita and the amount of tax per capita at the state level. Based on the following data, can she or he conclude that per capita state debt and per capita state taxes are related? Both amounts are in dollars and represent five randomly selected states.

Per capita debt x	1924	907	1445	1608	661
Per capita tax y	1685	1838	1734	1842	1317

Source: World Almanac

(The information in this exercise will be used for Exercises 16 and 37 in Section 10–2 and Exercises 18 and 22 in Section 10–3.)

17. Measles and Mumps A researcher wishes to see if there is a relationship between the number of reported cases of measles and mumps for a recent 5-year period. Is there a linear relationship between the two variables?

Measles cases	43	140	71	63	212
Mumps cases	800	454	1991	2612	370

Source: U.S. Department of Health

18. At Bats and Hits Is there a linear relationship between the number of hits a World Series player gets and the number of times at bat the player has?

At Bats	51	67	77	44	55	39	45
Hits	19	25	30	20	23	16	18

(The information in this exercise will be used for Exercise 18 in Section 10–2.)

19. Average Age and Length of Service Is the average age of armed services personnel related to the average amount of length of service in months? The data are shown.

Age	24.9	25.6	26.1	27.3	27
Time	66.5	70	74.8	89.6	82.6

(The information in this exercise will be used for Exercise 19 in Section 10–2.)

20. Water and Carbohydrates Here are the number of grams of water and the number of grams of carbohydrates for a random selection of raw foods (100 g each). Is there a linear relationship between the variables?

Water	83.93	80.76	87.66	85.20	72.85	84.61	83.81
Carbs	15.25	16.55	11.10	13.01	24.27	14.13	15.11

Source: Time Almanac.

(The information in this exercise will be used for Exercises 20 and 38 in Section 10–2.)

21. Faculty and Students The number of faculty and the number of students are shown for a random selection of small colleges. Is there a significant relationship between the two variables? Switch x and y and repeat the process. Which do you think is really the independent variable?

Faculty	99	110	113	116	138	174	220
Students	1353	1290	1091	1213	1384	1283	2075

Source: World Almanac.

(The information in this exercise will be used for Exercises 21 and 36 in Section 10–2.)

22. Life Expectancies Is there a relationship between the life expectancy for men and the life expectancy for women in a given country? A random sample of nonindustrialized countries was selected, and the life expectancy in years is listed for both men and women. Are the variables linearly related?

Men	59.7	72.9	41.9	46.2	50.3	43.2
Women	63.8	77.8	44.5	48.3	54.0	43.5

Source: World Almanac.

(The information in this exercise will be used for Exercise 22 in Section 10–2.)

23. Literacy Rates For the same countries used in Exercise 22, the literacy rates (in percents) for both men and women are listed. Is there a linear relationship

between the variables? (The information in this exercise will be used for Exercise 23 in Section 10–2.)

Men (%)	43.1	92.6	65.7	27.9	61.5	76.7
Women (%)	12.6	86.4	45.9	15.4	46.3	96.1

Source: World Almanac

24. NHL Assists and Total Points A random sample of scoring leaders from the NHL showed the following numbers of assists and total points. Based on these data, can it be concluded that there is a significant relationship between the two?

Assists	26	29	32	34	36	37	40
Total points	48	68	66	69	76	67	84

Source: Associated Press.

(The information in this exercise will be used for Exercise 24 in Section 10–2.)

25. Gestation and Average Longevity Is there a relationship between the gestation of a sample of animals and the average longevity in years for a random sample of animals?

Gestation x	105	285	151	238	112
Longevity y	5	15	8	41	10

(The information in this exercise will be used in Exercise 25 in Section 10–2.)

26. Tall Buildings An architect wants to determine the relationship between the heights (in feet) of buildings and the number of stories in the buildings. The data for a sample of 10 buildings in Chicago are shown. Explain the relationship (if any).

Stories x	64	68	50	48	32	46	58	45	49	40
Height y	995	844	732	679	648	635	610	600	583	573

(The information in this exercise will be used for Exercise 26 in Section 10–2.)

27. Class Size and Grades School administrators wondered whether class size and grade achievement (in percent) were related. A random sample of classes revealed the following data. Are the variables linearly related?

No. of students	15	10	8	20	18	6
Avg. grade (%)	85	90	82	80	84	92

(The information in this exercise will be used for Exercise 28 of this section and Exercise 27 in Section 10–2.)

Extending the Concepts

28. One of the formulas for computing r is

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{(n-1)(s_x)(s_y)}$$

Using the data in Exercise 27, compute r with this formula. Compare the results.

29. Compute r for the data set shown. Explain the reason for the value of r. Interchange the values of x and y. Compute r for this data set. Explain the results of the comparison.

x	1	2	3	4	5
y	4	7	10	13	16

30. Compute r for the following data and test the hypothesis H_0: $\rho = 0$. Draw the scatter plot; then explain the results.

x	−3	−2	−1	0	1	2	3
y	9	4	1	0	1	4	9

5–2 Regression

OBJECTIVE 4

Compute the equation of the regression line.

In studying relationships between two variables, collect the data and then construct a scatter plot. The purpose of the scatter plot, as indicated previously, is to determine the nature of the relationship between the variables. The possibilities include a positive linear relationship, a negative linear relationship, a curvilinear relationship, or no discernible relationship. After the scatter plot is drawn and a linear relationship is determined, the next steps are to compute the value of the correlation coefficient and to test the significance of the relationship. If the value of the correlation coefficient is significant, the next step is to determine the equation of the **regression line**, which is the data's

line of best fit. (*Note:* Determining the regression line when r is not significant and then making predictions using the regression line are meaningless.) The purpose of the regression line is to enable the researcher to see the trend and make predictions on the basis of the data.

Line of Best Fit

Historical Notes

Francis Galton drew the line of best fit visually. An assistant of Karl Pearson's named G. Yule devised the mathematical solution using the least-squares method, employing a mathematical technique developed by Adrien-Marie Legendre about 100 years earlier.

Figure 5–11 shows a scatter plot for the data of two variables. It shows that several lines can be drawn on the graph near the points. Given a scatter plot, you must be able to draw the *line of best fit. Best fit* means that the sum of the squares of the vertical distances from each point to the line is at a minimum.

The difference between the actual value y and the predicted value y' (that is, the vertical distance) is called a **residual** or a predicted error. Residuals are used to determine the line that best describes the relationship between the two variables.

The method used for making the residuals as small as possible is called the *method of least squares*. As a result of this method, the regression line is also called the *least squares regression line*.

The reason you need a line of best fit is that the values of y will be predicted from the values of x; hence, the closer the points are to the line, the better the fit and the prediction will be. See Figure 10–12. When r is positive, the line slopes upward and to the right. When r is negative, the line slopes downward from left to right.

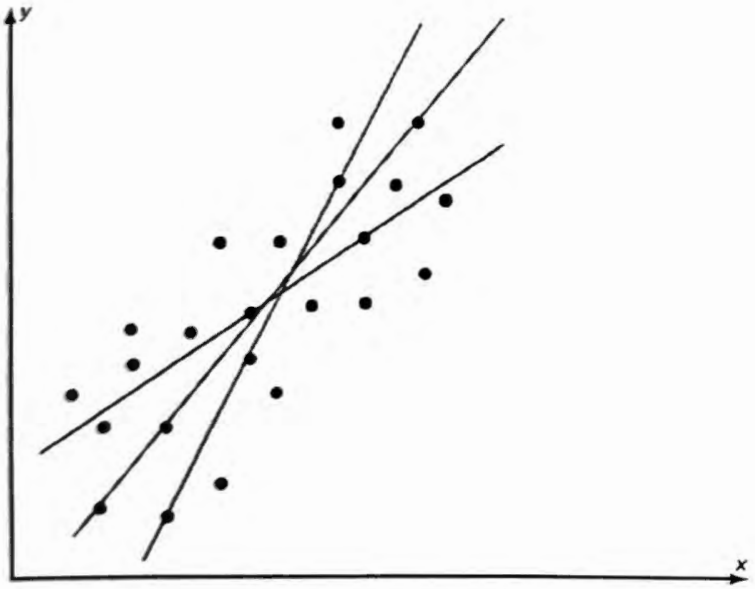


FIGURE 5–11
Scatter Plot with Three Lines Fit to the Data

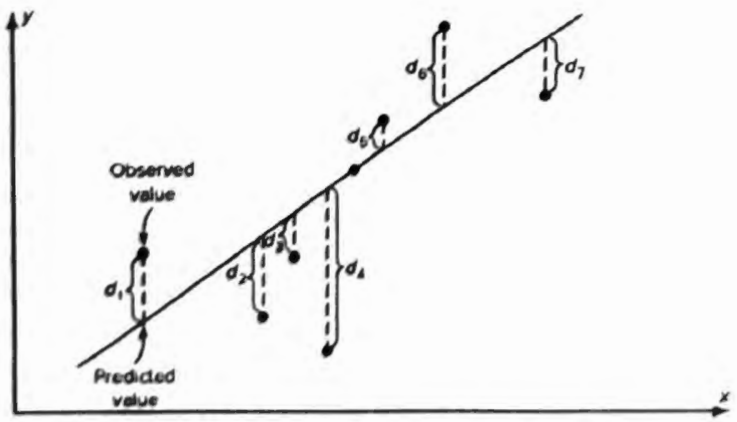
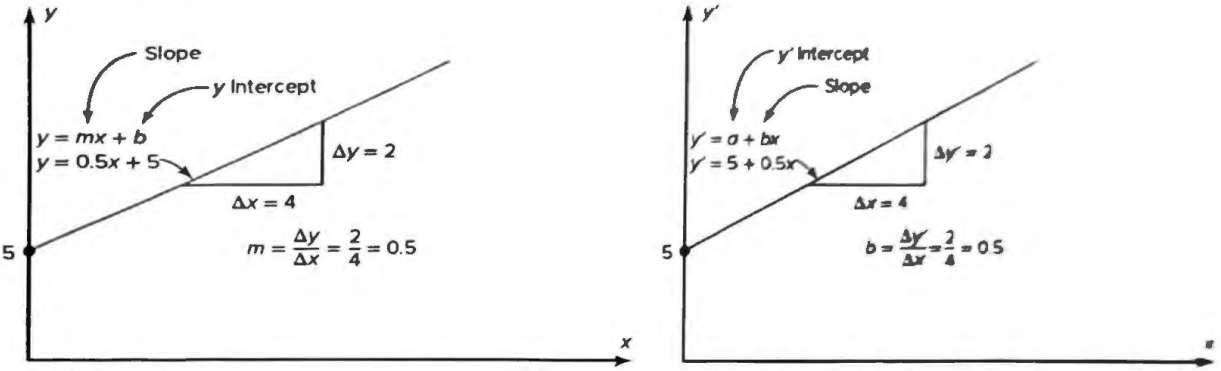


FIGURE 15–12
Line of Best Fit for a Set of Data Points

FIGURE 5–13 A Line as Represented in Algebra and in Statistics



(a) Algebra of a line

(b) Statistical notation for a regression line

Determination of the Regression Line Equation

In algebra, the equation of a line is usually given as $y = mx + b$, where m is the slope of the line and b is the y intercept. (Students who need an algebraic review of the properties of a line should refer to the online resources, before studying this section.) In statistics, the equation of the regression line is written as $y' = a + bx$, where a is the y' intercept and b is the slope of the line. See Figure 10–13.

There are several methods for finding the equation of the regression line. Two formulas are given here. *These formulas use the same values that are used in computing the value of the correlation coefficient.* The mathematical development of these formulas is beyond the scope of this book.

Formulas for the Regression Line $y' = a + bx$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

where a is the y' intercept and b is the slope of the line.

Rounding Rule for the Intercept and Slope Round the values of a and b to three decimal places.

The steps for finding the regression line equation are summarized in this Procedure Table.

Procedure Table

Finding the Regression Line Equation

Step 1 Make a table, as shown in step 2.

Step 2 Find the values of xy, x^2, and y^2. Place them in the appropriate columns and sum each column.

x	y	xy	x^2	y^2
$\Sigma x =$	$\Sigma y =$	$\Sigma xy =$	$\Sigma x^2 =$	$\Sigma y^2 =$

Step 3 When r is significant, substitute in the formulas to find the values of a and b for the regression line equation $y' = a + bx$.

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} \qquad b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

EXAMPLE 5–9 Car Rental Companies

Find the equation of the regression line for the data in Example 10–4, and graph the line on the scatter plot of the data.

SOLUTION

The values needed for the equation are $n = 6$, $\Sigma x = 153.8$, $\Sigma y = 18.7$, $\Sigma xy = 682.77$, and $\Sigma x^2 = 5859.26$. Substituting in the formulas, you get

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(18.7)(5859.26) - (153.8)(682.77)}{(6)(5859.26) - (153.8)^2} = 0.396$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{6(682.77) - (153.8)(18.7)}{(6)(5859.26) - (153.8)^2} = 0.106$$

Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 0.396 + 0.106x$$

To graph the line, select any two points for x and find the corresponding values for y. Use any x values between 10 and 60. For example, let $x = 15$. Substitute in the equation and find the corresponding y' value.

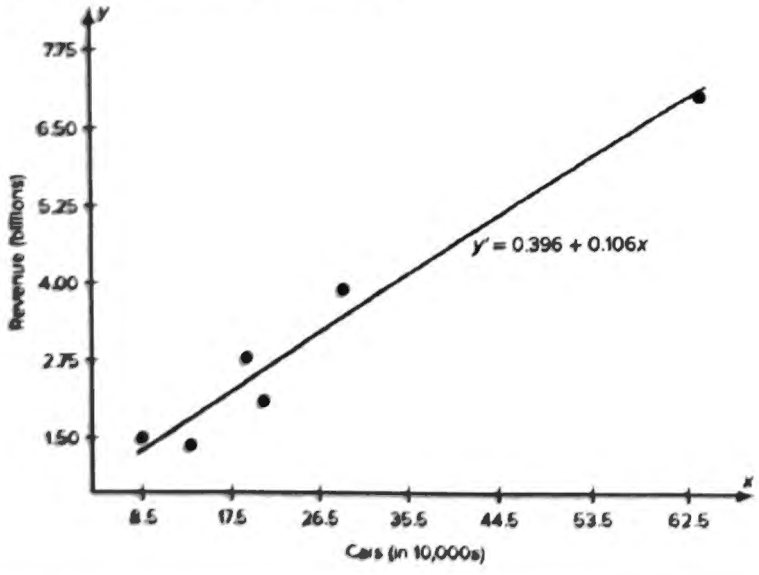
$$\begin{aligned} y' &= 0.396 + 0.106x \\ &= 0.396 + 0.106(15) \\ &= 1.986 \end{aligned}$$

Let $x = 40$; then

$$\begin{aligned} y' &= 0.396 + 0.106x \\ &= 0.396 + 0.106(40) \\ &= 4.636 \end{aligned}$$

Then plot the two points (15, 1.986) and (40, 4.636) and draw a line connecting the two points. See Figure 10–14.

FIGURE 5–14 Regression Line for Example 5–9



Note: When you draw the regression line, it is sometimes necessary to *truncate* the graph (see Chapter 2). This is done when the distance between the origin and the first labeled coordinate on the x axis is not the same as the distance between the rest of the labeled x coordinates or the distance between the origin and the first labeled y' coordinate is not the same as the distance between the other labeled y' coordinates. When the x axis or the y axis has been truncated, do not use the y' intercept value to graph the line. When you graph the regression line, always select x values between the smallest x data value and the largest x data value.

Historical Note

In 1795, Adrien-Marie Legendre (1752–1833) measured the meridian arc on the earth's surface from Barcelona, Spain, to Dunkirk, England. This measure was used as the basis for the measure of the meter. Legendre developed the least-squares method around the year 1805.

EXAMPLE 5–10 Absences and Final Grades

Find the equation of the regression line for the data in Example 5–5, and graph the line on the scatter plot.

SOLUTION

The values needed for the equation are $n = 7$, $\Sigma x = 57$, $\Sigma y = 511$, $\Sigma xy = 3745$, and $\Sigma x^2 = 579$. Substituting in the formulas, you get

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(511)(579) - (57)(3745)}{(7)(579) - (57)^2} = 102.493$$

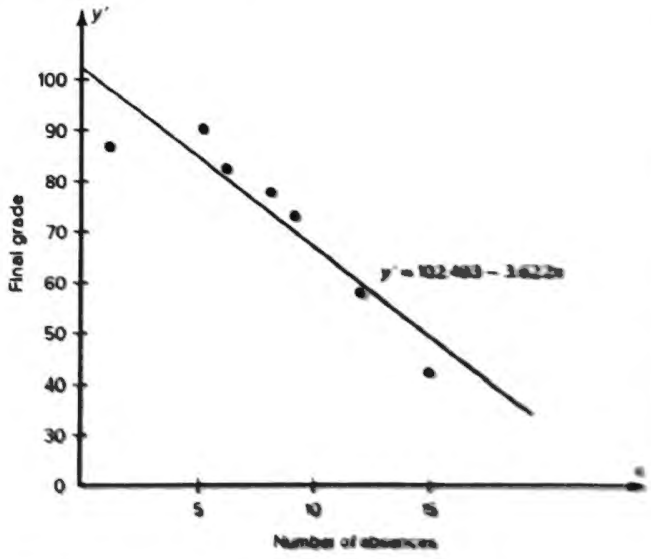
$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(7)(3745) - (57)(511)}{(7)(579) - (57)^2} = -3.622$$

Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 102.493 - 3.622x$$

The graph of the line is shown in Figure 10–15.

FIGURE 5–15
Regression Line for Example 5–10



The sign of the correlation coefficient and the sign of the slope of the regression line will always be the same. That is, if r is positive, then b will be positive; if r is negative, then b will be negative. The reason is that the numerators of the formulas are the same and determine the signs of r and b, and the denominators are always positive. The regression

line will always pass through the point whose x coordinate is the mean of the x values and whose y coordinate is the mean of the y values, that is, $(\bar{x}, \bar{y})$.

The regression line can be used to make predictions for the dependent variable. You should use these guidelines when you are making predictions.

1. The points of the scatter plot fit the linear regression line reasonably well.
2. The value of r is significant.
3. The value of a specific x is not much beyond the observed values (x values) in the original data.
4. If r is not significant, then the best predicted value for a specific x value is the mean of the y value in the original data.

Assumptions for Valid Predictions in Regression

1. The sample is a random sample.
2. For any specific value of the independent variable x, the value of the dependent variable y must be normally distributed about the regression line. See Figure 10–16(a).
3. The standard deviation of each of the dependent variables must be the same for each value of the independent variable. See Figure 10–16(b).

In this book, the assumptions will be stated in the exercises; however, when encountering statistics in other situations, you must check to see that these assumptions have been met before proceeding.

The method for making predictions is shown in Example 10–11.

EXAMPLE 5–11 Absences and Final Grades

Use the equation of the regression line in Example 5–10 to predict the final grade for a student who missed 4 classes.

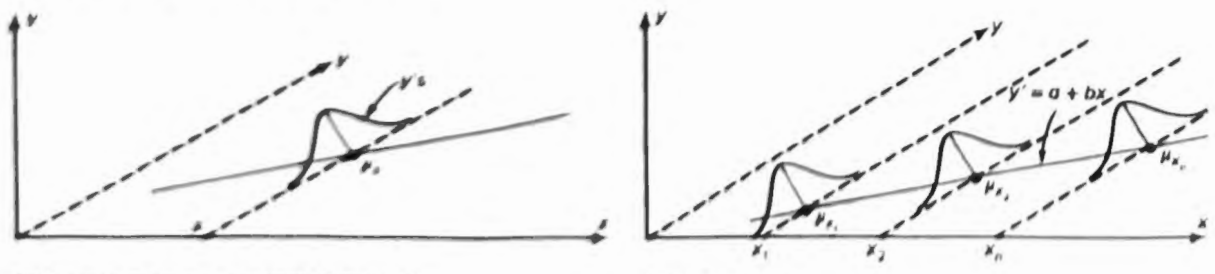
SOLUTION

Substitute 4 for x in the regression line equation $y' = 102.493 - 3.622x$.

$$\begin{aligned} y' &= 102.493 - 3.622x \\ &= 102.493 - 3.622(4) \\ &= 88.005 \\ &= 88 \text{ (rounded)} \end{aligned}$$

Hence, when a student misses 4 classes, the student's grade on the final exam is predicted to be about 88.

FIGURE 5–16 Assumptions for Predictions



(a) Dependent variable y normally distributed (b) $\sigma_1 = \sigma_2 = \cdots = \sigma_n$

The value obtained in Example 5–11 is a point prediction, and with point predictions, no degree of accuracy or confidence can be determined. More information on prediction is given in Section 10–3.

The magnitude of the change in one variable when the other variable changes exactly 1 unit is called a **marginal change.** The value of slope b of the regression line equation represents the marginal change. For example, in Example 10–9 the slope of the regression line is 0.106, which means for each additional increase of 10,000 cars, the value of y changes 0.106 unit ($106 million) on average.

Interesting Fact

It is estimated that wearing a motorcycle helmet reduces the risk of a fatal accident by 30%.

Extrapolation, or making predictions beyond the bounds of the data, must be interpreted cautiously. For example, in 1979, some experts predicted that the United States would run out of oil by the year 2003. This prediction was based on the current consumption and on known oil reserves at that time. However, since then, the automobile industry has produced many new fuel-efficient vehicles. Also, there are many as yet undiscovered oil fields. Finally, science may someday discover a way to run a car on something as unlikely but as common as peanut oil. In addition, the price of a gallon of gasoline was predicted to reach $10 a few years later. Fortunately this has not come to pass. *Remember that when predictions are made, they are based on present conditions or on the premise that present trends will continue.* This assumption may or may not prove true in the future.

A scatter plot should be checked for outliers. An outlier is a point that seems out of place when compared with the other points (see Chapter 3). Some of these points can affect the equation of the regression line. When this happens, the points are called **influential points** or **influential observations.**

© Malcolm Fife/age fotostock RF

When a point on the scatter plot appears to be an outlier, it should be checked to see if it is an influential point. An influential point tends to "pull" the regression line toward the point itself. To check for an influential point, the regression line should be graphed with the point included in the data set. Then a second regression line should be graphed that excludes the point from the data set. If the position of the second line is changed considerably, the point is said to be an influential point. Points that are outliers in the x direction tend to be influential points.

Researchers should use their judgment as to whether to include influential observations in the final analysis of the data. If the researcher feels that the observation is not necessary, then it should be excluded so that it does not influence the results of the study. However, if the researcher feels that it is necessary, then he or she may want to obtain additional data values whose x values are near the x value of the influential point and then include them in the study.

Applying the Concepts 5–2

Stopping Distances Revisited

In a study on speed and braking distance, researchers looked for a method to estimate how fast a person was traveling before an accident by measuring the length of the skid marks. An area that was focused on in the study was the distance required to completely stop a vehicle at various speeds. Use the following table to answer the questions.

MPH	Braking distance (feet)
20	20
30	45
40	81
50	133
60	205
80	411

Assume MPH is going to be used to predict stopping distance.

1. Find the linear regression equation.
2. What does the slope tell you about MPH and the braking distance? How about the y' intercept?
3. Find the braking distance when MPH = 45.
4. Find the braking distance when MPH = 100.
5. Comment on predicting beyond the given data values.

See page 604 for the answers.

Exercises 5–2

1. What two things should be done before one performs a regression analysis?
2. What are the assumptions for regression analysis?
3. What is the general form for the regression line used in statistics?
4. What is the symbol for the slope? For the y' intercept?
5. What is meant by the *line of best fit?*
6. When all the points fall on the regression line, what is the value of the correlation coefficient?
7. What is the relationship between the sign of the correlation coefficient and the sign of the slope of the regression line?
8. As the value of the correlation coefficient increases from 0 to 1, or decreases from 0 to −1, how do the points of the scatter plot fit the regression line?
9. How is the value of the correlation coefficient related to the accuracy of the predicted value for a specific value of x?
10. When the value of r is not significant, what value should be used to predict y?

For Exercises 11 through 27, use the same data as for the corresponding exercises in Section 10–1. For each exercise, find the equation of the regression line and find the y' value for the specified x value. Remember that no regression should be done when r is not significant.

11. **Crimes** The number of murders and robberies per 100,000 population for a random selection of states are shown.

Murders	2.4	2.7	5.6	2.6	2.1	3.3	6.6	5.7
Robberies	25.3	14.3	151.6	91.1	80	49	173	95.8

Find y' when $x = 4.5$ murders.

12. **Oil and Gas Prices** The average gasoline price per gallon (in cities) and the cost of a barrel of oil are shown below for a random selection of weeks in 2015.

Oil ($)	51.91	60.65	59.56	52.86	45.12	44.21
Gasoline ($)	1.97	1.96	2.06	2.04	2.00	1.99

Find the cost of gasoline when oil is $60 a barrel.

13. **Commercial Movie Releases** New movie releases per studio and gross receipts are as follows:

No. of releases	361	270	306	22	35	10	8	12	21
Gross receipts (million $)	3844	1962	1371	1064	334	241	188	154	125

Find y' when $x = 200$ new releases.

14. **Forest Fires and Acres Burned** Number of fires and number of acres burned are as follows:

Fires x	72	69	58	47	84	62	57	45
Acres y	62	42	19	26	51	15	30	15

Find y' when $x = 60$ fires.

15. **Alumni Contributions** Years and contribution data are as follows:

Years x	1	5	3	10	7	6
Contribution y, $	500	100	300	50	75	80

Find y' when $x = 4$ years.

16. **State Debt and Per Capita Taxes** Data for per capita state debt and per capita state tax are as follows:

Per capita debt	1924	907	1445	1608	661
Per capita tax	1685	1838	1734	1842	1317

Find y' when $x = \$1500$ in per capita debt.

17. Measles and Mumps The data show the number of cases of measles and mumps for a recent 5-year period.

Measles cases	43	140	71	63	212
Mumps cases	800	454	1991	2612	370

Given a year with 100 cases of measles, predict the expected number of cases of mumps for that year.

Source: U. S. Department of Health

18. At Bats and Hits The data show the number of hits and the number of at bats for 7 major league players in recent World Series.

At Bats	51	67	77	44	55	39	45
Hits	19	25	30	20	23	16	18

Find y' when $x = 60$.

19. Average Age and Length of Service The data show the average ages and lengths of service in months.

Age	24.9	25.6	26.1	27.3	27
Time	66.5	70	74.8	89.6	82.6

If a service person is 26.8 years old, predict the time that the person will serve.

20. Water and Carbohydrates Here are the number of grams of water and the number of grams of carbohydrates for a random selection of raw foods (100 g each).

Water	83.93	80.76	87.66	85.20	72.85	84.61	83.81
Carbs	15.25	16.55	11.10	13.01	24.27	14.13	15.11

Find y' for $x = 75$.

21. Faculty and Students The number of faculty and the number of students in a random selection of small colleges are shown.

Faculty	99	110	113	116	138	174	220
Students	1353	1290	1091	1213	1384	1283	2075

Now find the equation of the regression line when x is the variable of the number of students.

22. Life Expectancies A random sample of nonindustrialized countries was selected, and the life expectancy in years is listed for both men and women.

Men	59.7	72.9	41.9	46.2	50.3	43.2
Women	63.8	77.8	44.5	48.3	54.0	43.5

Find women's life expectancy in a country where men's life expectancy = 60 years.

23. Literacy Rates For the same countries used in Exercise 22, the literacy rates (in percents) for both men and women are listed.

Men	43.1	92.6	65.7	27.9	61.5	76.7
Women	12.6	86.4	45.9	15.4	46.3	96.1

Find y' when $x = 80$.

24. NHL Assists and Total Points The number of assists and the total number of points for a sample of NHL scoring leaders are shown.

Assists	26	29	32	34	36	37	40
Total points	48	68	66	69	76	67	84

Find y' when $x = 30$ assists.

25. Gestation and Average Longevity The data show the gestation period in days and the longevity of the lifetime of the animals in years. Predict y' if $x = 200$ days.

Gestation x	105	285	151	238	112
Longevity y	5	15	8	41	10

26. Tall Buildings The data show the number of stories and the heights (in feet) of a sample of buildings in Chicago.

Stories x	64	68	50	48	32	46	58	45	49	40
Height y	995	844	732	679	648	635	610	600	583	573

For a building with 55 stories, how high would we expect it be?

27. Class Size and Grades School administrators wondered whether class size and grade achievement (in percent) were related. A random sample of classes revealed the following data.

No. of students	15	10	8	20	18	6
Avg. grade (%)	85	90	82	80	84	92

Find y' when $x = 12$.

For Exercises 28 through 33, do a complete regression analysis by performing these steps.

a. Draw a scatter plot.
b. Compute the correlation coefficient.
c. State the hypotheses.
d. Test the hypotheses at $\alpha = 0.05$. Use Table I.
e. Determine the regression line equation if r is significant.
f. Plot the regression line on the scatter plot, if appropriate.
g. Summarize the results.

28. Fireworks and Injuries These data were obtained for the years 2009 through 2014 and indicate the number of fireworks (in millions) used and the related injuries.

Predict the number of injuries if 100 million fireworks are used during a given year.

Fireworks in use x	213.9	205.9	234.1	207.5	186.4	225.3
Related injuries y	8800	8600	9600	8600	11,400	10,500

Source: National Council of Fireworks Safety, American Pyrotechnic Assoc.

29. Farm Acreage Is there a relationship between the number of farms in a state and the acreage per farm? A random selection of states across the country, both eastern and western, produced the following results. Can a relationship between these two variables be concluded?

No. of farms (thousands) x	77	52	20.8	49	28	58.2
Acreage per farm y	347	173	173	218	246	132

Source: *World Almanac.*

30. SAT Scores Educational researchers desired to find out if a relationship exists between the average SAT verbal score and the average SAT mathematical score. Several states were randomly selected, and their SAT average scores are recorded below. Is there sufficient evidence to conclude a relationship between the two scores?

Verbal x	526	504	594	585	503	589
Math y	530	522	606	588	517	589

Source: *World Almanac.*

31. Coal Production These data were obtained from a sample of counties in southwestern Pennsylvania and indicate the number (in thousands) of tons of bituminous coal produced in each county and the number of employees working in coal production in each county. Predict the amount of coal produced for a county that has 500 employees.

No. of employees x	110	731	1031	20	118	1162	103	752
Tons y	227	5410	5328	147	729	8095	635	6157

32. Television Viewers A television executive selects 10 television shows and compares the average number of viewers the show had last year with the average number of viewers this year. The data (in millions) are shown. Describe the relationship.

Viewers last year x	26.6	17.85	20.3	16.8	20.8
Viewers this year y	28.9	19.2	26.4	13.7	20.2
Viewers last year x	16.7	19.1	18.9	16.0	15.8
Viewers this year y	18.8	25.0	21.0	16.8	15.3

Source: Nielsen Media Research.

33. Absences and Final Grades An educator wants to see how the number of absences for a student in her class affects the student's final grade. The data obtained from a sample are shown.

No. of absences x	10	12	2	0	8	5
Final grade y	70	65	96	94	75	82

For Exercises 34 and 35, do a complete regression analysis and test the significance of r at $\alpha = 0.05$, using the P-value method.

34. Father's and Son's Weights A physician wishes to know whether there is a relationship between a father's weight (in pounds) and his newborn son's weight (in pounds). The data are given here.

Father's weight x	176	160	187	210	196	142	205	215
Son's weight y	6.6	8.2	9.2	7.1	8.8	9.3	7.4	8.6

35. Age and Net Worth Is a person's age related to his or her net worth? A sample of 10 billionaires is selected, and the person's age and net worth are compared. The data are given here.

Age x	56	39	42	60	84	37	68	66	73	55
Net worth y (billion $)	18	14	12	14	11	10	10	7	7	5

Source: The Associated Press.

Extending the Concepts

36. For Exercises 13, 15, and 21 in Section 5–1, find the mean of the x and y variables. Then substitute the mean of the x variable into the corresponding regression line equations found in Exercises 13, 15, and 21 in this section and find y'. Compare the value of y' with $\bar{y}$ for each exercise. Generalize the results.

37. The y intercept value a can also be found by using the equation

$$a = \bar{y} - b\bar{x}$$

Verify this result by using the data in Exercises 15 and 16 of Sections 5–1 and 5–2.

38. The value of the correlation coefficient can also be found by using the formula

$$r = \frac{bs_x}{s_y}$$

where s_x is the standard deviation of the x values and s_y is the standard deviation of the y values. Verify this result for Exercises 18 and 20 of Section 5–1.

Data Display

Row	Subject	Absences	Final Grade	FITS	RESI
1	A	6	82	80.7612	1.23881
2	B	2	86	95.2488	−9.24876
3	C	15	43	48.1642	−5.16418
4	D	9	74	69.8955	4.10448
5	E	12	58	59.0299	−1.02985
6	F	5	90	84.3831	5.61692
7	G	8	78	73.5174	4.48259

5–3 Coefficient of Determination and Standard Error of the Estimate

The previous sections stated that if the correlation coefficient is significant, the equation of the regression line can be determined. Also, for various values of the independent variable x, the corresponding values of the dependent variable y can be predicted. Several other measures are associated with the correlation and regression techniques. They include the coefficient of determination, the standard error of the estimate, and the prediction interval. But before these concepts can be explained, the different types of variation associated with the regression model must be defined.

Types of Variation for the Regression Model

Consider the following hypothetical regression model.

	1	2	3	4	5
	10	8	12	16	20

The equation of the regression line is $y' = 4.8 + 2.8x$, and $r = 0.919$. The sample y values are 10, 8, 12, 16, and 20. The predicted values, designated by y', for each x can be found by substituting each x value into the regression equation and finding y'. For example, when $x = 1$,

$$y' = 4.8 + 2.8x = 4.8 + (2.8)(1) = 7.6$$

Now, for each x, there is an observed y value and a predicted y' value; for example, when $x = 1$, $y = 10$ and $y' = 7.6$. Recall that the closer the observed values are to the predicted values, the better the fit is and the closer r is to $+1$ or -1.

The *total variation* $\Sigma(y - \bar{y})^2$ is the sum of the squares of the vertical distances each point is from the mean. The total variation can be divided into two parts: that which is attributed to the relationship of x and y and that which is due to chance. The variation obtained from the relationship (i.e., from the predicted y' values) is $\Sigma(y' - \bar{y})^2$ and is called the *explained variation.*

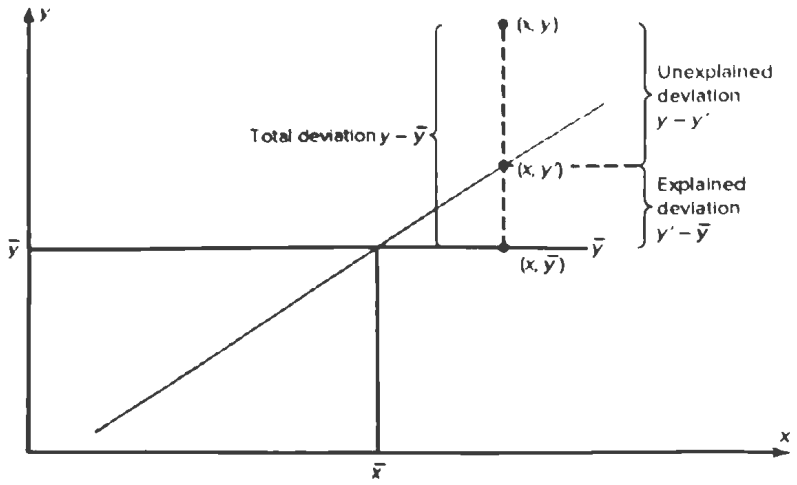
In other words, the explained variation is the vertical distance $y' - \bar{y}$, which is the distance between the predicted value y' and the mean value $\bar{y}$. Most of the variation can be explained by the relationship. The closer the value r is to $+1$ or -1, the better the points fit the line and the closer $\Sigma(y' - \bar{y})^2$ is to $\Sigma(y - \bar{y})^2$. In fact, if all points fall on the regression line, $\Sigma(y' - \bar{y})^2$ will equal $\Sigma(y - \bar{y})^2$, since y' is equal to y in each case.

On the other hand, the variation due to chance, found by $\Sigma(y - y')^2$, is called the *unexplained variation.* In other words, the unexplained variation is the vertical distance $y - y'$, which is the distance between the observed value, y, and the predicted value y'. This variation cannot be attributed to the relationship. When the unexplained variation is small, the value of r is close to $+1$ or -1. If all points fall on the regression line, the unexplained variation $\Sigma(y - y')^2$ will be 0. Hence, the *total variation* is equal to the sum of the explained variation and the unexplained variation. That is,

$$\Sigma(y - \bar{y})^2 = \Sigma(y' - \bar{y})^2 + \Sigma(y - y')^2$$

These values are shown in Figure 10–17. For a single point, the differences are called *deviations.* For the hypothetical regression model given earlier, for $x = 1$ and $y = 10$, you get $y' = 7.6$ and $\bar{y} = 13.2$.

FIGURE 5-17
Deviations for the Regression Equation



The procedure for finding the three types of variation is illustrated next.

Step 1 Find the predicted y' values.

For $x = 1$ $y' = 4.8 + 2.8x = 4.8 + (2.8)(1) = 7.6$

For $x = 2$ $y' = 4.8 + (2.8)(2) = 10.4$

For $x = 3$ $y' = 4.8 + (2.8)(3) = 13.2$

For $x = 4$ $y' = 4.8 + (2.8)(4) = 16.0$

For $x = 5$ $y' = 4.8 + (2.8)(5) = 18.8$

Unusual Stat

There are 1,929,770,126,028,800 different color combinations for Rubik's cube and only one correct solution in which all the colors of the squares on each face are the same.

Hence, the values for this example are as follows:

x	y	y'
1	10	7.6
2	8	10.4
3	12	13.2
4	16	16.0
5	20	18.8

Step 2 Find the mean of the y values.

$$\bar{y} = \frac{10 + 8 + 12 + 16 + 20}{5} = 13.2$$

Step 3 Find the total variation $\Sigma(y - \bar{y})^2$.

$(10 - 13.2)^2 = 10.24$

$(8 - 13.2)^2 = 27.04$

$(12 - 13.2)^2 = 1.44$

$(16 - 13.2)^2 = 7.84$

$(20 - 13.2)^2 = 46.24$

$\Sigma(y - \bar{y})^2 = 92.8$

Step 4 Find the explained variation $\Sigma(y' - \bar{y})^2$.

$(7.6 - 13.2)^2 = 31.36$

$(10.4 - 13.2)^2 = 7.84$

$(13.2 - 13.2)^2 = 0.00$

$(16 - 13.2)^2 = 7.84$

$(18.8 - 13.2)^2 = 31.36$

$\Sigma(y' - \bar{y})^2 = 78.4$

Historical Note

In the 19th century, astronomers such as Gauss and Laplace used what is called the principle of least squares based on measurement errors to determine the shape of Earth. It is now used in regression theory.

Step 5 Find the unexplained variation $\Sigma(y - y')^2$.

$$(10 - 7.6)^2 = 5.76$$
$$(8 - 10.4)^2 = 5.76$$
$$(12 - 13.2)^2 = 1.44$$
$$(16 - 16)^2 = 0.00$$
$$(20 - 18.8)^2 = 1.44$$
$$\Sigma(y - y')^2 = 14.4$$

Notice that

Total variation = explained variation + unexplained variation

$$92.8 = 78.4 + 14.4$$

Residual Plots

As previously stated, the values $y - y'$ are called *residuals* (sometimes called the *prediction errors*). These values can be plotted with the x values, and the plot, called a **residual plot,** can be used to determine how well the regression line can be used to make predictions.

The residuals for the previous example are calculated as shown.

1	10	7.6	$10 - 7.6 = 2.4$
2	8	10.4	$8 - 10.4 = -2.4$
3	12	13.2	$12 - 13.2 = -1.2$
4	16	16	$16 - 16 = 0$
5	20	18.8	$20 - 18.8 = 1.2$

The x values are plotted using the horizontal axis, and the residuals are plotted using the vertical axis. Since the mean of the residuals is always zero, a horizontal line with a y coordinate of zero is placed on the y axis as shown in Figure 10–18.

Plot the x and residual values as shown in Figure 10–18.

	1	2	3	4	5
	2.4	−2.4	−1.2	0	1.2

To interpret a residual plot, you need to determine if the residuals form a pattern. Figure 10–19 shows four examples of residual plots. If the residual values are more or less evenly distributed about the line, as shown in Figure 10–19(a), then the relationship between x and y is linear and the regression line can be used to make predictions.

FIGURE 5–18
Residual Plot

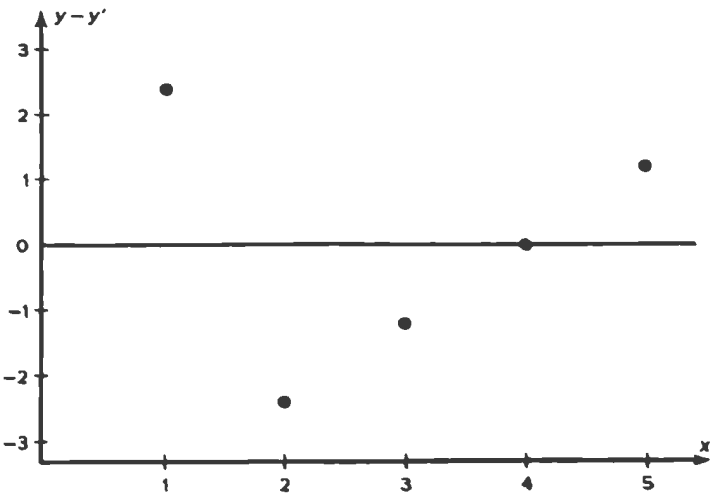
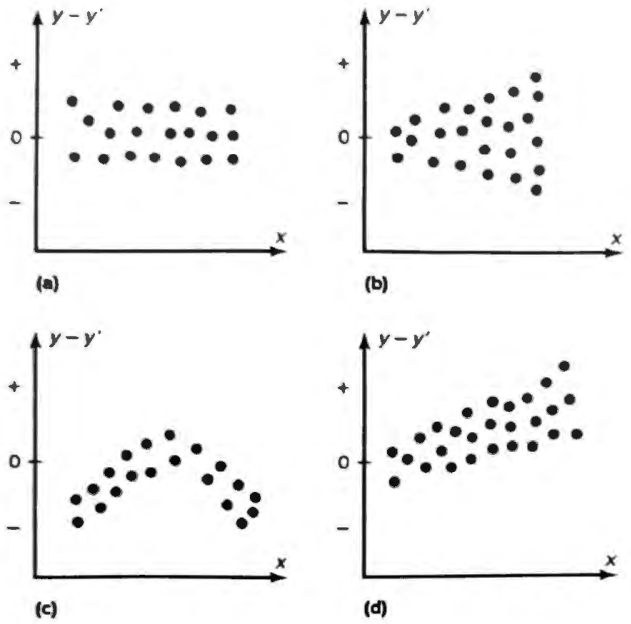


FIGURE 5–19
Examples of Residual Plots



This means that the standard deviations of each of the dependent variables must be the same for each value of the independent variable. This is called the *homoscedasticity assumption.* See assumption 3 on page 568.

Figure 10–19(b) shows that the variance of the residuals increases as the values of x increase. This means that the regression line is not suitable for predictions.

Figure 10–19(c) shows a curvilinear relationship between the x values and the residual values; hence, the regression line is not suitable for making predictions.

Figure 10–19(d) shows that as the x values increase, the residuals increase and become more dispersed. This means that the regression line is not suitable for making predictions.

The residual plot in Figure 10–18 shows that the regression line $y' = 4.8 + 2.8x$ is somewhat questionable for making predictions due to a small sample size.

OBJECTIVE 5

Compute the coefficient of determination.

Coefficient of Determination

The *coefficient of determination* is the ratio of the explained variation to the total variation and is denoted by r^2. That is,

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

The coefficient of determination is a measure of the variation of the dependent variable that is explained by the regression line and the independent variable. The symbol for the coefficient of determination is r^2.

Historical Note

Karl Pearson recommended in 1897 that the French government close all its casinos and turn the gambling devices over to the academic community to use in the study of probability

The coefficient of determination is a number between 0 and 1 inclusive, or $0 \leq r^2 \leq 1$. If $r^2 = 0$, then the least squares regression line cannot explain any of the variation. If $r^2 = 1$, the least squares regression line explains 100% of the variation in the dependent variable.

For the example, $r^2 = 78.4/92.8 = 0.845$. The term r^2 is usually expressed as a percentage. So in this case, 84.5% of the total variation is explained by the regression line using the independent variable.

Another way to arrive at the value for r^2 is to square the correlation coefficient. In this case, $r = 0.919$ and $r^2 = 0.845$, which is the same value found by using the variation ratio.

Of course, it is usually easier to find the coefficient of determination by squaring r and converting it to a percentage. Therefore, if $r = 0.90$, then $r^2 = 0.81$, which is equivalent to 81%. This result means that 81% of the variation in the dependent variable is accounted for by the variation in the independent variable. The rest of the variation, 0.19, or 19%, is unexplained. This value is called the *coefficient of nondetermination* and is found by subtracting the coefficient of determination from 1. As the value of r approaches 0, r^2 decreases more rapidly. For example, if $r = 0.6$, then $r^2 = 0.36$, which means that only 36% of the variation in the dependent variable can be attributed to the variation in the independent variable.

Coefficient of Nondetermination

$$1.00 - r^2$$

In Example 10–5 using absences and final grades, the correlation coefficient is $r = -0.944$. Then $r^2 = (-0.944)^2 = 0.891$. Hence, about 0.891, or 89.1%, of the variation in the final grades can be explained by the linear relationship between the number of absences and the final grades. About $1 - 0.891$ or 0.109 or 10.9% of the variation of the final grades cannot be explained by the variation of the absences.

Standard Error of the Estimate

OBJECTIVE 6

Compute the standard error of the estimate.

When a y' value is predicted for a specific x value, the prediction is a point prediction. The disadvantage of a point prediction is that it doesn't give us any information about how accurate the point prediction is. In previous chapters, we developed confidence interval estimates to overcome this disadvantage. In this section, we will use what is called a **prediction interval,** which is an interval estimate of a variable.

Recall in previous chapters an interval estimate of a parameter, such as the mean or standard deviation, is called a confidence interval.

A prediction interval is an interval estimate of a predicted value of y when the regression equation is used and a specific value of x is given.

A prediction interval about the y' value can be constructed, just as a confidence interval was constructed for an estimate of the population mean. The prediction interval uses a statistic called the *standard error of the estimate.*

The standard error of the estimate, denoted by s_{est}, is the standard deviation of the observed y values about the predicted y' values. The formula for the standard error of the estimate is

$$s_{est} = \sqrt{\frac{\Sigma(y - y')^2}{n - 2}}$$

The standard error of the estimate is similar to the standard deviation, but the mean is not used. Recall that the standard deviation measures how the values deviate from the mean. The standard error of the estimate measures how the data points deviate from the regression line.

As can be seen from the formula, the standard error of the estimate is the square root of the unexplained variation—that is, the variation due to the difference of the observed values and the expected values—divided by $n - 2$. So the closer the observed values are to the predicted values, the smaller the standard error of the estimate will be.

A Procedure Table for finding the standard error of the estimate is shown here.

Procedure Table

Finding the Standard Error of the Estimate

Step 1 Make a table using the column headings shown.

x	y	y'	$y-y'$	$(y-y')^2$

Step 2 Find the predicted values y' for each x value, and place these values under y' in the table.

Step 3 Subtract each y' value from each y value, and place these answers in the $y - y'$ column in the table.

Step 4 Square each of the values in step 3, and place these values in the column $(y - y')^2$.

Step 5 Find the sum of the values in the $(y - y')^2$ column.

Step 6 Substitute in the formula and find s_{est}.

$$s_{est} = \sqrt{\frac{\Sigma(y-y')^2}{n-2}}$$

Example 10–12 shows how to compute the standard error of the estimate.

EXAMPLE 10–12 Copy Machine Maintenance Costs

A researcher collects the following data and determines that there is a significant relationship between the age of a copy machine and its monthly maintenance cost. The regression equation is $y' = 55.57 + 8.13x$. Find the standard error of the estimate.

Machine	Age x (years)	Monthly cost y
A	1	$ 62
B	2	78
C	3	70
D	4	90
E	4	93
F	6	103

SOLUTION

Step 1 Make a table, as shown.

x	y	y'	$y-y'$	$(y-y')^2$
1	62			
2	78			
3	70			
4	90			
4	93			
6	103			

Step 2 Using the regression line equation $y' = 55.57 + 8.13x$, compute the predicted values y' for each x, and place the results in the column labeled y'.

$x = 1$ $y' = 55.57 + (8.13)(1) = 63.70$
$x = 2$ $y' = 55.57 + (8.13)(2) = 71.83$
$x = 3$ $y' = 55.57 + (8.13)(3) = 79.96$
$x = 4$ $y' = 55.57 + (8.13)(4) = 88.09$
$x = 6$ $y' = 55.57 + (8.13)(6) = 104.35$

Step 3 For each y, subtract y' and place the answer in the column labeled $y - y'$.

$$62 - 63.70 = -1.70 \qquad 90 - 88.09 = 1.91$$
$$78 - 71.83 = 6.17 \qquad 93 - 88.09 = 4.91$$
$$70 - 79.96 = -9.96 \qquad 103 - 104.35 = -1.35$$

Step 4 Square the numbers found in step 3 and place the squares in the column labeled $(y - y')^2$.

Step 5 Find the sum of the numbers in the last column. The completed table is shown.

x	y	y'	$y-y'$	$(y-y')^2$
1	62	63.70	−1.70	2.89
2	78	71.83	6.17	38.0689
3	70	79.96	−9.96	99.2016
4	90	88.09	1.91	3.6481
4	93	88.09	4.91	24.1081
6	103	104.35	−1.35	1.8225
			$\Sigma(y-y')^2 = 169.7392$	

Step 6 Substitute in the formula and find s_{est}.

$$s_{est} = \sqrt{\frac{\Sigma(y - y')^2}{n - 2}} = \sqrt{\frac{169.7392}{6 - 2}} = 6.514$$

In this case, the standard deviation of observed values about the predicted values is 6.514.

The standard error of the estimate can also be found by using the formula

$$s_{est} = \sqrt{\frac{\Sigma y^2 - a\,\Sigma y - b\,\Sigma xy}{n - 2}}$$

This Procedure Table shows the alternate method for finding the standard error of the estimate.

Procedure Table

Alternative Method for Finding the Standard Error of Estimate

Step 1 Make a table using the column headings shown.

x	y	xy	y^2

Step 2 Place the x values in the first column (the x column), and place the y values in the second column (the y column). Find the products of the x and y values and place them in the third column (the xy column). Square the y values and place them in the fourth column (the y^2 column).

Step 3 Find the sum of the values in the y, xy, and y^2 columns.

Step 4 Identify a and b from the regression equation, substitute in the formula, and evaluate.

$$s_{est} = \sqrt{\frac{\Sigma y^2 - a\,\Sigma y - b\,\Sigma xy}{n - 2}}$$

EXAMPLE 5–13

Find the standard error of the estimate for the data for Example 10–12 by using the preceding formula. The equation of the regression line is $y' = 55.57 + 8.13x$.

SOLUTION

Step 1 Make a table as shown in the Procedure Table.

Step 2 Place the x values in the first column (the x column), and place the y values in the second column (the y column). Find the product of x and y values, and place the results in the third column. Square the y values, and place the results in the y^2 column.

Step 3 Find the sums of the y, xy, and y^2 columns. The completed table is shown here.

1	62	62	3,844
2	78	156	6,084
3	70	210	4,900
4	90	360	8,100
4	93	372	8,649
6	103	618	10,609
	$\Sigma y = 496$	$\Sigma xy = 1778$	$\Sigma y^2 = 42,186$

Step 4 From the regression equation $y' = 55.57 + 8.13x$, $a = 55.57$, and $b = 8.13$. Substitute in the formula and solve for s_{est}.

$$s_{est} = \sqrt{\frac{\Sigma y^2 - a\,\Sigma y - b\,\Sigma xy}{n-2}}$$

$$= \sqrt{\frac{42,186 - (55.57)(496) - (8.13)(1778)}{6-2}} = 6.483$$

This value is close to the value found in Example 10–12. The difference is due to rounding.

OBJECTIVE 7

Find a prediction interval.

Prediction Interval

The standard error of the estimate can be used for constructing a prediction interval (similar to a confidence interval) about a y' value.

Recall that when a specific value x is substituted into the regression equation, the predicted value y' that you get is a point estimate for y. For example, if the regression line equation for the age of a machine and the monthly maintenance cost is $y' = 55.57 + 8.13x$ (Example 10–12), then the predicted maintenance cost for a 3-year-old machine would be $y' = 55.57 + 8.13(3)$, or \$79.96. Since this is a point estimate obtained from the regression equation, you have no idea how accurate it is because there are possible sources of prediction errors in finding the regression line equation. One source occurs when finding the standard error of the estimate s_{est}. Two others are errors made in estimating the slope and the y' intercept, since the equation of the regression line will change somewhat if different random samples are used when calculating the equation. However, you can construct a prediction interval about the estimate. By selecting an α value, you can achieve $(1 - \alpha) \cdot 100\%$ confidence that the interval contains the actual mean of the y values that correspond to the given value of x.

Formula for the Prediction Interval about a Value y'

$$y' - t_{\alpha/2}s_{est}\sqrt{1+\frac{1}{n}+\frac{n(x-\overline{X})^2}{n\Sigma x^2-(\Sigma x)^2}} < y < y' + t_{\alpha/2}s_{est}\sqrt{1+\frac{1}{n}+\frac{n(x-\overline{X})^2}{n\Sigma x^2-(\Sigma x)^2}}$$

with d.f. $= n-2$.

The next Procedure Table can be used to find the prediction.

Procedure Table

Finding a Prediction Interval for a Specific Independent Data Value

Step 1 Find Σx, Σx^2, and $\overline{x}$.

Step 2 Find y' for the specific x value.

Step 3 Find s_{est}.

Step 4 Substitute in the formula and evaluate.

$$y' - t_{\alpha/2}s_{est}\sqrt{1+\frac{1}{n}+\frac{n(x-\overline{X})^2}{n\Sigma x^2-(\Sigma x)^2}} < y < y' + t_{\alpha/2}s_{est}\sqrt{1+\frac{1}{n}+\frac{n(x-\overline{X})^2}{n\Sigma x^2-(\Sigma x)^2}}$$

with d.f. $= n-2$.

EXAMPLE 5–14

For the data in Example 10–12, find the 95% prediction interval for the monthly maintenance cost of a machine that is 3 years old.

SOLUTION

Step 1 Find Σx, Σx^2, and $\overline{X}$.

$$\Sigma x = 20 \qquad \Sigma x^2 = 82 \qquad \overline{X} = \frac{20}{6} = 3.3$$

Step 2 Find y' for $x = 3$.

$$y' = 55.57 + 8.13x$$
$$= 55.57 + 8.13(3) = 79.96$$

Step 3 Find s_{est}.

$$s_{est} = 6.48$$

as shown in Example 10–13.

Step 4 Substitute in the formula and solve: $t_{\alpha/2} = 2.776$, d.f. $= 6 - 2 = 4$ for 95%.

$$y' - t_{\alpha/2}s_{est}\sqrt{1+\frac{1}{n}+\frac{n(x-\overline{X})^2}{n\Sigma x^2-(\Sigma x)^2}} < y < y' + t_{\alpha/2}s_{est}\sqrt{1+\frac{1}{n}+\frac{n(x-\overline{X})^2}{n\Sigma x^2-(\Sigma x)^2}}$$

$$79.96 - (2.776)(6.48)\sqrt{1+\frac{1}{6}+\frac{6(3-3.3)^2}{6(82)-(20)^2}} < y < 79.96 + (2.776)(6.48)\sqrt{1+\frac{1}{6}+\frac{6(3-3.3)^2}{6(82)-(20)^2}}$$

$$79.96 - (2.776)(6.48)(1.08) < y < 79.96 + (2.776)(6.48)(1.08)$$
$$79.96 - 19.43 < y < 79.96 + 19.43$$
$$60.53 < y < 99.39$$

Hence, you can be 95% confident that the interval $60.53 < y < 99.39$ contains the actual value of y.

That is, if a copy machine is 3 years old, we can be 95% confident that the maintenance cost would be between \$60.53 and \$99.39. This range is large because the sample size is small, $n = 6$, and the standard error of estimate is 6.48.

Applying the Concepts 5–3

Interpreting Simple Linear Regression

Answer the questions about the following computer-generated information.

Linear correlation coefficient $r = 0.794556$
Coefficient of determination = 0.631319
Standard error of estimate = 12.9668
Explained variation = 5182.41
Unexplained variation = 3026.49
Total variation = 8208.90
Equation of regression line $y' = 0.725983x + 16.5523$
Level of significance = 0.1
Test statistic = 0.794556
Critical value = 0.378419

1. Are both variables moving in the same direction?
2. Which number measures the distances from the prediction line to the actual values?
3. Which number is the slope of the regression line?
4. Which number is the y intercept of the regression line?
5. Which number can be found in a table?
6. Which number is the allowable risk of making a type I error?
7. Which number measures the variation explained by the regression?
8. Which number measures the scatter of points about the regression line?
9. What is the null hypothesis?
10. Which number is compared to the critical value to see if the null hypothesis should be rejected?
11. Should the null hypothesis be rejected?

See page 604 for the answers.

Exercises 5–3

1. What is meant by the *explained variation*? How is it computed?
2. What is meant by the *unexplained variation*? How is it computed?
3. What is meant by the *total variation*? How is it computed?
4. Define the coefficient of determination.
5. How is the coefficient of determination found?
6. Define the coefficient of nondetermination.
7. How is the coefficient of nondetermination found?

For Exercises 8 through 13, find the coefficients of determination and nondetermination and explain the meaning of each.

8. $r = 0.62$
9. $r = 0.44$
10. $r = 0.51$
11. $r = 0.97$

12. $r = 0.12$

13. $r = 0.15$

14. Define the standard error of the estimate for regression. When can the standard error of the estimate be used to construct a prediction interval about a value y'?

15. Compute the standard error of the estimate for Exercise 13 in Section 10–1. The regression line equation was found in Exercise 13 in Section 10–2.

16. Compute the standard error of the estimate for Exercise 14 in Section 10–1. The regression line equation was found in Exercise 14 in Section 10–2.

17. Compute the standard error of the estimate for Exercise 15 in Section 10–1. The regression line equation was found in Exercise 15 in Section 10–2.

18. Compute the standard error of the estimate for Exercise 16 in Section 10–1. The regression line equation was found in Exercise 16 in Section 10–2.

19. For the data in Exercises 13 in Sections 10–1 and 10–2 and 15 in Section 10–3, find the 90% prediction interval when $x = 200$ new releases.

20. For the data in Exercises 14 in Sections 10–1 and 10–2 and 16 in Section 10–3, find the 95% prediction interval when $x = 60$.

21. For the data in Exercises 15 in Sections 10–1 and 10–2 and 17 in Section 10–3, find the 90% prediction interval when $x = 4$ years.

22. For the data in Exercises 16 in Sections 10–1 and 10–2 and 18 in Section 10–3, find the 98% prediction interval when $x = 47$ years.

5–4 Multiple Regression (Optional)

OBJECTIVE 8

Be familiar with the concept of multiple regression.

The previous sections explained the concepts of simple linear regression and correlation. In simple linear regression, the regression equation contains one independent variable x and one dependent variable y' and is written as

$$y' = a + bx$$

where a is the y' intercept and b is the slope of the regression line.

In **multiple regression,** there are several independent variables and one dependent variable, and the equation is

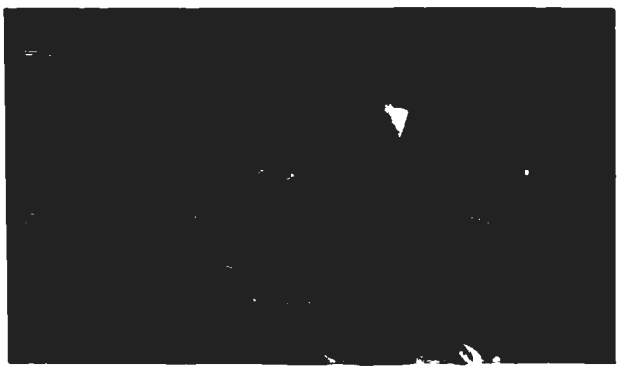
$$y' = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

where $x_1, x_2, \ldots, x_k$ are the independent variables.

For example, suppose a nursing instructor wishes to see whether there is a relationship between a student's grade point average, age, and score on the state board nursing examination. The two independent variables are GPA (denoted by x_1) and age (denoted by x_2). The instructor will collect the data for all three variables for a sample of nursing students. Rather than conduct two separate simple regression studies, one using the GPA

Unusual Stats

The most popular single-digit number played by people who purchase lottery tickets is 7



© Norma Jean Gargasz/Alamy